Estudos das propriedades da Inosina em DNA através do modelo Peyrard-Bishop e análise dos parâmetros termodinâmicos utilizados na predição de estruturas secundárias de RNA

Rodolfo Vieira Maximiano

Dezembro 2017

Universidade Federal de Minas Gerais - UFMG

Instituto de Ciências Exatas - ICEx

Programa de Pós Graduação em Física

Estudos das propriedades da Inosina em DNA através do modelo Peyrard-Bishop e análise dos parâmetros termodinâmicos utilizados na predição de estruturas secundárias de RNA

Rodolfo Vieira Maximiano

Orientador: Prof. Gerald Weber

Tese apresentada ao Programa de Pós-Graduação em Física do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Física.

Dezembro 2017

Agradecimentos

Dedico esta tese à minha família, especialmente minha mãe. Vocês sempre me deram todo apoio para perseguir meus sonhos. Mãe, Tata, e Jonathan: obrigado por serem parte da minha vida.

Agradeço ao professor Gerald Weber que tanto me ensinou durante todo o período que trabalhamos juntos. Discussões que não se restrigiam apenas à vida acadêmica e nossa atuação profissional como físicos, professores, e pesquisadores, mas que também traziam ensinamentos sobre sabedoria, generosidade e parcimônia. Agradeço o apoio e a paciência, tão grandes que eu não tentarei descrever.

Agradeço ainda à todos que me cobriram de estímulo, carinho, companheirismo, risadas, compreensão, paciência, durante todo o processo de doutoramento no departamento de física. Em especial agradeço à Tauanne, Luciana, Marco, Marcos, Erik, Iza, Vivi, dentre tantos outros que não poderia citar de forma compreensiva.

Agradeço à cada um dos professores que me serviram de exemplo (de didática, de comprometimento, de paixão pela física). Gostaria de citar, especialmente, o outros orientadores que tive. Agradeço ao prof. Newton Martins Barbosa pela orientação de iniciação científica que me fez desejar ser um cientista e acreditar no Brasil. Agradeço ao prof. Luiz Gustavo Cançado por ter me recebido em Belo Horizonte, e por todos diálogos durante o mestrado.

Also, I would like to thank professor David Mathews from University of Rochester Medical Center, for kindly providing us with information about experimentally determined secondary structures of many kinds of RNAs. E agradeço também ao pessoal da UFRGS, em especial Ana Luiza Maia e Lucieli Ceolin, pela colaboração que nos permitiu desenvolver o trabalho conjunto de análise das sequências gênicas RET.

Finalmente, agradeço às agências de fomento que foram a razão pela qual pude fazer minha pós-graduação. Agradeço ao CNPq, FAPEMIG e CAPES. Em especial, agradeço à CAPES pela minha bolsa de doutorado, e à FAPEMIG pelo fomento ao grupo de pesquisa do qual fiz parte.

Resumo

Nesta tese apresentaremos diferentes estudos sobre as propriedades de ácidos nucleicos, tanto em termos estruturais, quanto em sua composição. Após uma breve introdução nos conceitos de ácido desoxirribonucleico (*deoxyribonucleic acid*, DNA), ácido ribonucléico (*ribonucleic acid*, RNA), e sobre o nucleotídeo da inosina, apresentamos os trabalhos realizados. Em nosso primeiro trabalho fazemos a análise termodinâmica das propriedades físicas dos pareamentos da inosina com diferentes tipos de base que ocorrem no DNA. Usando dados experimentais fomos capazes de determinar os parâmetros físicos do modelo Peyrard-Bishop (PB) que descrevem as configurações das ligações de hidrogênio da inosina com cada base do DNA, e os que descrevem as interações de empilhamento quando o pareamento que contém inosina está adjacente a outros pares do tipo Watson-Crick ao longo da fita. Nossos resultados indicam que duas ligações de hidrogênio são formadas em todos pareamentos de inosina, o que também é uma indicação verificada em estudos de dinâmica molecular envolvendo a hipoxantina — a nucleobase da inosina. Além disso, pudemos encontrar que as interações de empilhamento da inosina com diferentes tipos de pares adjacentes são comparáveis aos dos pares canônicos entre si, com a exceção da adjacência da inosina-guanina, que aparentemente distorce a dupla hélice de DNA de tal forma que torna o empilhamento muito forte ou muito fraco dependendo de qual base irá se acoplar à ele. O segundo trabalho apresentado é um estudo sobre a estabilidade de sequências de RNA através da análise da estrutura secundária das mesmas, e análise da influência da incerteza experimental das medidas dos parâmetros termodinâmicos usados no modelo de próximos vizinhos (*nearest-neighbour model*, NN) na qualidade da predição de softwares de uso comum para a determinação estruturas secundárias de sequências de RNA. Usando engenharia reversa fomos capazes de utilizar dados de NMR e difração de raio X para determinar o melhor conjunto de parâmetros que otimizam a predição de estruturas. Encontramos que para sequências longas até 800 nucleotídios o uso dos *softwares* de predição não garante uma predição realista. Milhares de estruturas variantes foram encontradas, que também são igualmente válidas nos limites do erro experimental dos parâmetros de próximos vizinhos usados como argumentos em tais programas, mas que diferem enormemente da estrutura medida experimentalmente. Portanto, esperamos que nossa investigação possa atrair a atenção para os limites apropriados de utilização dos *softwares* de predição de estrutura secundária.

Palavras-chaves: RNA,DNA, modelo Peyrard-Bishop, Estrutura secundária de RNA

Abstract

In this thesis we present different studies on the properties of nucleic acids regarding both structure and their composition. After an brief overview on *deoxyribonucleic acid* (DNA), *ribonucleic acid* (RNA), and the inosine nucleotide, given in the introduction, we start to present our results. In the first part we have the thermodynamic analysis of the physical properties of inosine pairing occurring when present in DNA. Using experimental data we were able to determine the Peyrard-Bishop parameters that describe its hydrogen bonding configurations with each type of DNA base, and the stacking interactions between inosine mismatches and regular neighbour Watson-Crick pairs. Our findings indicate two hydrogen bonds in all situations, which was also indicated in molecular dynamics studies of oligomers of hypoxantine — the inosine nucleobase. In addition, we found that the stacking interactions between inosine mismatches are comparable with canonical interactions, with the exception of IG bases which seem to distort the DNA double helix to cause either a very strong or very weak coupling. The second work we present is a study on the stability of RNA sequences through analysis of their secondary structures, and the analysis of the influence of experimental deviation uncertainty of nearest-neighbour model parameters measurements on the prediction quality of common softwares used for secondary structure determination of RNA sequences. Using reverse engineering we were able to use secondary structure data from NMR and X-ray crystallography to determine the best set of parameters to optimize secondary structure prediction. We have found that for longer sequences, up to 800 nucleotides, the software does not guarantee a realistic prediction. We found thousands of structural variations are equally valid within the experimental thermodynamic parameters used as arguments in for these software packages. Yet most of these structures differ from the experimentally determined structures. Therefore, we hope that our study may draw the attention to the proper limits for the use of these secondary structure software packages.

Keywords: RNA, DNA, Peyrard-Bishop Model, RNA secondary structure english

Sumário

Resumo **I**

Abstract **II**

1 Introdução **1**
1.1 Contexto e motivação 1
1.2 Organização da tese 2
1.3 A estrutura do DNA 3
1.4 A estrutura do ácido ribonucléico (*ribonucleic acid*, RNA) 6
1.5 Conformações dos pareamentos 8
1.6 Conclusão 8

2 Termodinâmica de experimentos de desnaturação **9**
2.1 Experimento de desnaturação de DNA 9
2.1.1 Desnaturação por dois estados 10
2.2 Modelo de próximos vizinhos 12

3 Modelo Peyrard-Bishop **15**
3.1 Introdução 15
3.2 O modelo Peyrard-Bishop para ácido desoxirribonucleico (*deoxyribonucleic acid*, DNA) homogêneo 16
3.2.1 O método da matriz de transferência 18
3.2.2 Estiramento médio entre as bases 19
3.2.3 O problema da divergência 20
3.3 O modelo Peyrard-Bishop para ácido desoxirribonucleico (*deoxyribonucleic acid*, DNA) heterogêneo 21
3.3.1 Método da equivalência térmica 22
3.3.2 Procedimento de cálculo de desnaturação 23
3.4 Parametrização do modelo PB 25
3.5 Conclusão 26

4 Inosina em DNA pelo modelo Peyrard-Bishop (PB) **27**
4.1 A estrutura da Inosina 27
4.2 Motivação e objetivo 29
4.3 Dados de desnaturação 29
4.4 Procedimento de minimização 30
4.5 Resultados para os parâmetros de Morse 31
4.6 Resultados para os parâmetros de empilhamento 32
4.7 Análise dos perfis de abertura de sequências contendo inosina 34
4.8 Conclusões 35

5 Estrutra e funcionalidades do ácido ribonucléico (*ribonucleic acid*, RNA) 36
5.1 rRNA — RNA ribossomal . . . 36
5.2 tRNA — RNA de transferência . . . 37
5.3 mRNA — RNA mensageiro . . . 39
5.4 miRNA, siRNA — RNAs que regulam outros RNAs . . . 40
5.5 Conclusão . . . 41

6 Estrutura Secundária de RNA 42
6.1 Motivação . . . 42
6.2 Elementos das estruturas secundárias . . . 43
6.2.1 Representações da estrutura secundária . . . 45
6.3 Algoritmos para predição da estrutura secundária . . . 48
6.3.1 Softwares de predição de estrutura secundária . . . 51
6.4 Estruturas secundárias subóptimas . . . 52
6.4.1 Estruturas subóptimas de mRNA de MTC . . . 56
6.5 Influência dos parâmetros experimentais na predição de estruturas secundárias . . . 60
6.5.1 Procedimento . . . 61
6.5.2 Otimização dos parâmetros termodinâmicos de estruturas secundárias . . . 65
6.5.3 Índices da qualidade de predição . . . 65
6.5.4 Procedimento de otimização . . . 65
6.5.5 Investigação da melhora nos índices de qualidade . . . 67
6.6 Conclusão . . . 68

Conclusões 69

A Minha trajetória profissional durante o doutorado 71

B Códigos Perl utilizados 73
B.1 Average-repetitions-per-size.pl . . . 73
B.2 repeated-structures.pl . . . 74
B.3 separating-results.pl . . . 75
B.4 calculador-ppv-sensitividade.pl . . . 76
B.5 ppv-sensitivity.pl . . . 79
B.5.1 ppv-sensitivity-step1.pl . . . 79
B.5.2 ppv-sensitivity-step2.pl . . . 80
B.5.3 ppv-sensitivity-step3.pl . . . 84
B.6 randomize-parameters.pl . . . 84
B.6.1 all-together-randomize-gaussian-parameters.pl . . . 85

C Melhores Parâmetros termodinâmicos 91

D Artigo publicado inosina em DNA pelo modelo Peyrard-Bishop 93

E Artigo publicado no estudo de caso do câncer de tireóide MCT-MEN2-RET 99

Referências Bibliográficas 115

Abreviações usadas

CT tabela de conectividade (*connectivity table*) pp. 47, 48

DFT teoria do funcional de densidade (*density functional theory*) pp. 15, 29, 32

DNA ácido desoxirribonucleico (*deoxyribonucleic acid*) pp. 3–17, 21–29, 32–35, 39, 69

MFEEU MFE com incerteza experimental (*MFE within experimental uncertainty*) pp. 60–64, 68, 69

LSU sub-unidade maior (*large sub-unit*) pp. 36, 37

MD dinâmica molecular (*molecular dynamics*) p. 15

MFE energia livre mínima (*minimum free energy*) pp. 51, 52, 57–60, 63

miRNA microRNA (*microRNA*) pp. 36, 40, 41

mRNA RNA mensageiro (*messenger RNA*) pp. 36, 38–42, 56–58, 60, 68

MTC carcinoma medular da tireoide (*medullary thyroid carcinoma*) pp. 42, 56

NMR ressonância magnética nuclear (*nuclear magnetic ressonance*) pp. 5, 28, 70

NN modelo de próximos vizinhos (*nearest-neighbour model*) pp. 9, 12, 13, 41, 60

PB modelo Peyrard-Bishop pp. 2, 15, 16, 18, 26, 27, 29, 30

PCR reação em cadeia da polimerase (*polymerase chain reaction*) p. 28

PPV valor positivo predito (*positive predicted value*) pp. 65–68

RET rearranjado durante a transfecção (*REarranged during Transfection*) pp. 56–58

RNA ácido ribonucléico (*ribonucleic acid*) pp. 2, 4, 6–8, 14, 15, 25, 36, 39–49, 51, 57, 62, 65, 68

rRNA RNA ribossomal (*ribosomal RNA*) pp. 36, 37

siRNA RNA curto de interferência (*small interfering RNA*) pp. 36, 40

SNP polimorfismo de nucleotídeo único (*single nucleotide polimorfism*) pp. 43, 56, 57

SSU sub-unidade menor (*small sub-unit*) pp. 36, 37

tRNA RNA de transferência (*transfer RNA*) pp. 27, 36–39, 52

3′UTR região 3′ não-traduzida (*3′ untranslated region*) p. 57

UV ultra-violeta pp. 9, 10, 28

WT selvagem (*wild-type*) pp. 56–59

Capítulo 1

Introdução

1.1 Contexto e motivação

Não há maneira melhor de descrever-se a motivação em se estudar ácidos nucleicos do que mencionar o Projeto Genoma [1] e a sua ambição inicial, que podem ser vistas no seguinte trecho extraído da página oficial do programa:

> The Human Genome Project (HGP) was the international, collaborative research program whose goal was the complete mapping and understanding of all the genes of human beings. All our genes together are known as our "genome".
>
> The HGP was the natural culmination of the history of genetics research. In 1911, Alfred Sturtevant, then an undergraduate researcher (...) realized that he could - and had to, in order to manage his data - map the locations of (...) genes whose mutations the Morgan laboratory was tracking over generations. Sturtevant's very first gene map can be likened to the Wright brothers' first flight at Kitty Hawk. In turn, the Human Genome Project can be compared to the Apollo program bringing humanity to the moon.

Este texto resume bem a ambição, em um contexto geral, de se entender os ácidos nucleicos, suas propriedades, comportamentos e funcionalidades. Não apenas o DNA da espécie humana, que é obviamente de grande interesse para a comunidade científica, mas também o de outras espécies. Bioquímica, física, biólogos, e diversos outras áreas contribuem para os desenvolvimentos feitos na pesquisa dessa área, e estes esforços conjuntos são, indubitavelmente, fonte de grandes avanços para o nosso entendimento sobre o tema.

Nesta tese de física demos grande importância às informações estruturais do DNA e RNA. Nos preocupamos em entender, por exemplo, como a conformação estrutural da molécula se altera quando ocorrem substituições de bases canônicas por não-canônicas na sequência. De maneira geral, nosso grupo busca estudar os ácidos nucleicos utilizando modelos mesoscópicos em conjunto com dados experimentais fornecidos por outras técnicas (como NMR e difração de raio X)[2].

A pesquisa em genética não é apenas útil para entender a natureza que nos cerca no presente momento: a promessa dos sequenciadores genéticos portáteis disponíveis nos próximos anos, e o desenvolvimento de técnicas poderosas de edição gênica como a CRISPR/CAS9[3], faz com que a discussão sobre este tema seja necessária para o entendimento de nosso futuro próximo. Especula-se que poderemos eliminar doenças genéticas de nossas populações, que conseguiremos realizar uma medicina personalizada e editar organismos geneticamente modificados para melhor eficiência na agricultura[4].

1.2 Organização da tese

Neste capítulo trarei uma visão geral sobre o que será discutido ao longo da tese, iniciando com uma seção motivacional sobre os motivos que nos levam a estudar ácidos nucleicos. Em seguida teremos, ainda neste capítulo introdutório, algumas informações sobre as moléculas biológicas de DNA e RNA, com as quais trabalharemos ao longo deste texto. Além disso, discutiremos a presença de inosina em DNA, e sua capacidade de se parear com todas outras bases.

No capítulo 2 iremos abordar a teoria dos processos termodinâmicos de desnaturação de DNA. Esta parte merece uma atenção especial pois diversas outras discussões serão fundamentadas nos resultados que tais experimentos nos fornecem. Além dos experimentos em si, discutiremos o modelo de próximos vizinhos, ao qual também nos referimos recorrentemente ao longo deste texto.

O capítulo 3 apresentará o modelo Peyrard-Bishop para análise do DNA e ácido ribonucléico (*ribonucleic acid*, RNA). Como discutiremos, este é um modelo mesoscópico que nos permite obter diversas informações sobre a estrutura molecular das fitas e dos pareamentos entre as bases. Este foi o modelo com o qual escolhemos abordar os pareamentos de inosina.

O capítulo 4, logo, mostra quais foram nossos resultados ao investigarmos a influência da inosina nas sequências de DNA. Utilizando o modelo Peyrard-Bishop, e os experimentos de desnaturação das fitas, conseguimos descrever características estruturais sobre as interações entre a inosina, seu

paramento, e os pares de base vizinhos.

O capítulo 5 traz o tema das estruturas secundárias. Definiremos o que são as estruturas secundárias, discutiremos como funcionam os algoritmos de predição das mesmas — o que inclui dados experimentais dos estudos discutidos no capítulo dois. Veremos como as incertezas nos experimentos de desnaturação influencia na habilidade do algoritmo de chegar em estruturas resposta que possuam relevância física.

Por fim, no capítulo 6.6, antes dos apêndices, teremos as conclusões sobre cada um dos estudos realizados.

1.3 A estrutura do DNA

Iniciaremos com uma introdução do que é conhecido sobre as moléculas de ácido desoxirribonucleico (*deoxyribonucleic acid*, DNA), e está bem consolidado na literatura [5, 6]. Todas as deslumbrantes capacidades desta molécula na manutenção da vida não serão discutidas nesta tese, onde focaremos nos aspectos físicos que nos são de interesse para a tentativa de compreendê-la cientificamente.

A molécula de DNA é um ácido nucleico composto por uma cadeia de açúcar-fosfato acoplado a diversas bases nitrogenadas. Seu nome, ácido desoxirribonucleico, vem justamente do tipo de açúcar que compõe tal cadeia. Em várias situações esta cadeia é chamada de esqueleto de açúcar-fosfato ou *scaffold* (“andaime”). A cadeia de açúcar possui uma orientação particular que é identificada pela disposição dos átomos de carbono que pertencem aos anéis da mesma: dizemos que a orientação padrão é do átomo de carbono $5'$ para o $3'$. As bases nitrogenadas são adenina, citosina, guanina, timina (A, C, G, T, respectivamente). Estes quatro tipos de base são as componentes com as quais é escrito o código que dará todas informações necessárias para desenvolvimento completo de um indivíduo de uma dada espécie. Em algumas espécies o DNA fica livre pelo citoplasma, em outras é armazenado no núcleo (para células que possuem tal separação física).

O DNA possui fitas duplas, também chamadas de hélices duplas, que se ligam de forma complementar através de pontes de hidrogênio formadas entre as bases nitrogenadas pertencentes à cada fita. Usualmente, A liga-se com T, C liga-se com G. As figuras 1.1 e 1.2 mostram como são os pareamentos tradicionais destas bases, que recebem o nome de pareamentos Watson-Crick, em homenagem aos cientistas James Watson e Francis Crick ganhadores do prêmio Nobel por determina-

rem a estrutura do DNA. O pareamento CG é feito por três ligações de hidrogênio como mostra a figura 1.1. O paramento AT, por sua vez, é feito com duas ligações de hidrogênio, veja a figura 1.2.

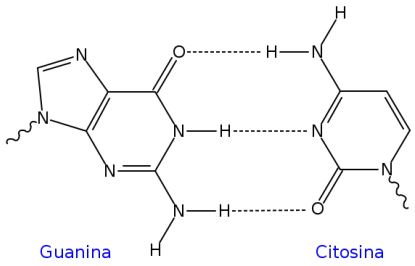


Figura 1.1
Pareamento entre citosina e guanina no DNA. Três pares de ligações de hidrogênio são formadas. Figura da referência 7.

A hélice dupla do DNA tem algumas características próprias: durante a torção em formato de hélice a sua estrutura final termina com sítios mais curtos ou mais longos entre um passo da volta e outro. O sítio maior é chamado de *major groove* e possui cerca de 10 Å de comprimento, enquanto o sítio menor, *minor groove*, possui cerca de 5 Å de comprimento. A figura 1.3 ilustra estes sítios. Além disso, a distância entre pares de base vizinhos é de cerca de 3.4 Å, este valor é chamado na literatura de *rise step*. Estes parâmetros podem variar de acordo com cada tipo de hélice, como veremos adiante.

Outras informações físicas relevantes seguem: o DNA por si só não teria estabilidade pois ambas as fitas possuem densidades de carga negativas, se repelindo. Elas conseguem se manter conectadas, naturalmente, pela presença de íons positivos a seu redor (por exemplo, íons de Na^{+} ou Mg^{2+}) [10]. O interior das fitas é extremamente hidrofóbico, com pouquíssima água adentrando o interior da fita onde ocorre o pareamento das bases. Sabe-se ainda que o DNA forma mais de um tipo de hélice dupla (e em alguns casos até hélices triplas [11, 12] ou quádruplas [13, 14]). Alguns exemplos dos tipos diferentes de hélices estão na figura 1.4. A hélice do tipo A é preferida pelo RNA, o DNA em condições fisiológicas faz hélices do tipo B, os tipos C e Z são quando temos baixa umidade com presença de íons Li^{+} ou Mg^{2+}, e alta concentração salina ou neutralização de cargas, respectivamente.

A interação de uma fita com a outra não é a única presente. Ao longo da mesma fita temos inte-

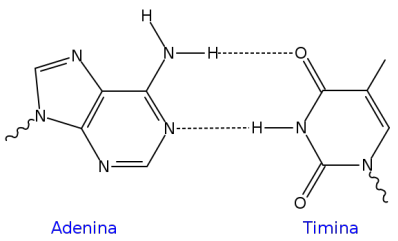


Figura 1.2
Pareamento entre Adenina e Timina no DNA. Dois pares de ligação de hidrogênio são formadas. Figura da referência 8.

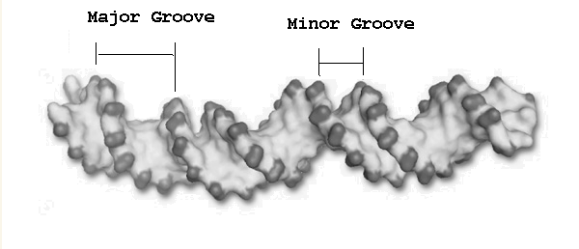


Figura 1.3
Ilustração dos sítios maior (*major groove*) e menor (*minor groove*) de uma dupla hélice de DNA. Figura da referência 9.

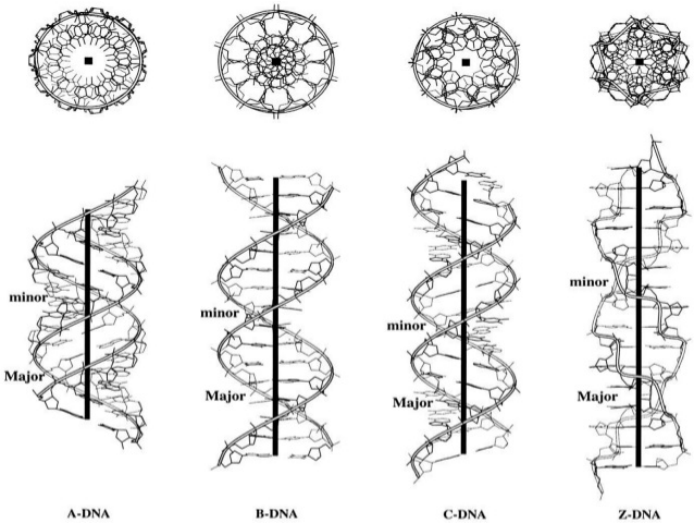


Figura 1.4
Ilustração dos tipos de hélice passíveis de serem formados pelos ácidos nucleicos. Figura da referência 15.

rações que chamamos de interações de empilhamento, *stacking*, e que surgem devido à interação de Van-der-Walls, interações dipolo-dipolo induzido, hidrofobia das fitas e à sobreposição dos orbitais pi de uma base nitrogenada com as suas vizinhas [2]. Além disso temos evidências que sugerem contribuições de forças de dispersão de London [2]. Estas forças de dispersão são proporcionais ao produto dos valores de polaridade e são inversamente proporcionais à sexta potência da distância que separa as bases. A figura 1.5 ilustra estes empilhamentos bem como o direcionamento anti-paralelo das duas fitas, isto é, uma delas na direção 5′-3′ e a outra na direção 3′-5′. Em condições especiais as duas fitas também podem parear em conformação paralela,

Sobre os empilhamentos temos algumas características específicas obtidas através de ressonân-

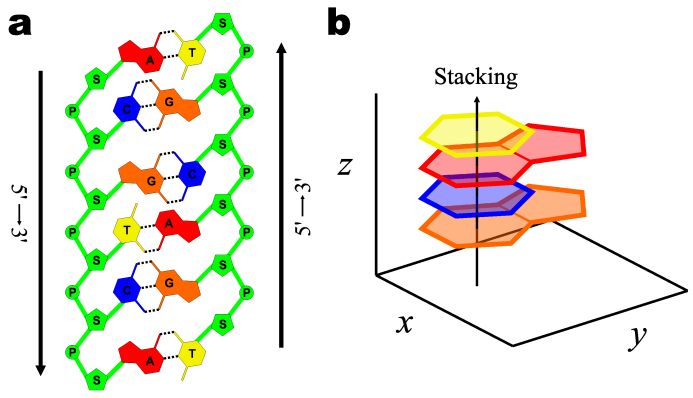


Figura 1.5
Ilustração dos tipos de ligação em uma hélice dupla e do direcionamento 5′-3′. Em (a) temos ligações covalentes (linhas contínuas) e ligações de hidrogênio (linhas pontilhadas). Em (b) vemos o empilhamento de quatro bases adjacentes. Figura da referência 16.

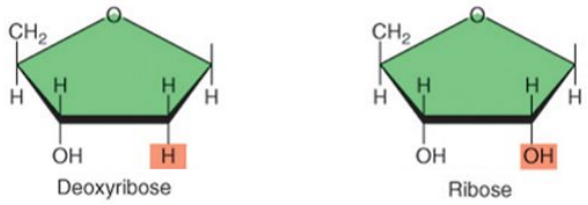


Figura 1.6
Desoxirribose (esquerda) em comparação com a ribose (direta). Ambos são açúcares que compõe a cadeia açúcar-fosfato. Figura da referência 17.

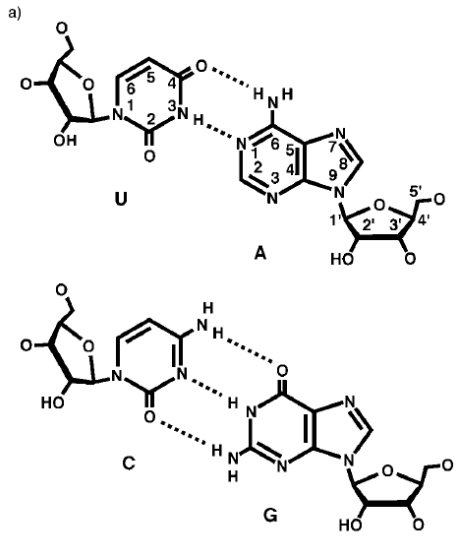


Figura 1.7
Pareamento tipo Watson-Crick em RNA, uracila-adenina (acima) e citosina-guanosina (em baixo). Figura da referência 18.

cia magnética nuclear (*nuclear magnetic ressonance*, NMR) em meio aquoso [2]. Por tais experimentos sabemos, por exemplo, que os processos que agregam as bases em interações de empilhamento, e posteriormente as desfazem são rápidos o suficiente para serem controlados por processos de difusão. As interações de empilhamento vão além da formação de estados do tipo dímero: elas entram no escopo das chamadas interações isodésmicas — as adições de mais bases ao conjunto empilhado é aditiva e não-cooperativa, e cada passo independente mostra os mesmos valores de parâmetros termodinâmicos e de cinética química [2]. Quanto aos tipos de empilhamento, experimentos de solubilidade mostraram que o empilhamento entre purinas é mais estável. A ordem de estabilidade é como segue: purina-purina $>$ pirimidina-purina $>$ pirimidina-pirimidina [2].

1.4 A estrutura do RNA

Embora o RNA possua a mesma estrutura básica do DNA, cadeia de açúcar-fosfato acompanhada de bases nitrogenadas, as diferenças entre as moléculas são muitas [2], e discutiremos várias delas nesta introdução. O açúcar que compõe a estrutura é a ribose, em contraste com a desoxirribose mencionada anteriormente para DNA, veja a figura 1.6. A outra diferença importante é a presença de uracila (U) no lugar da timina. A estrutura de fita dupla existe e em geral assume uma conforma-

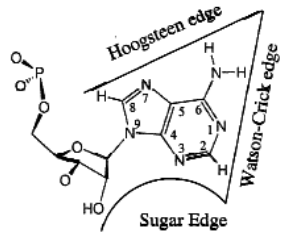
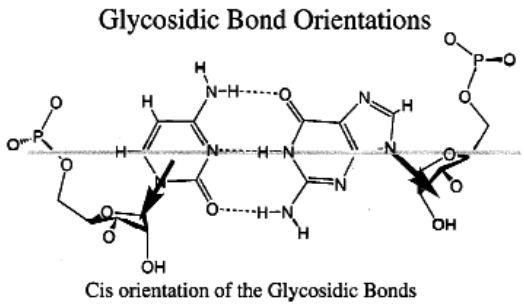


Figura 1.8
As faces possíveis de pareamento em RNA. Figura da referência 20.



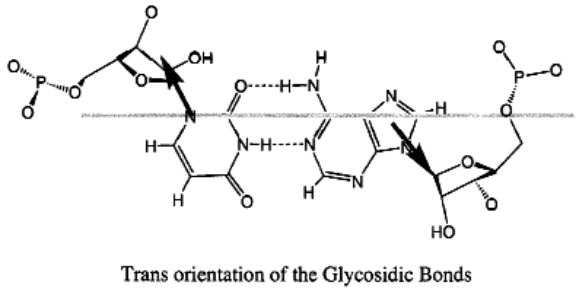


Figura 1.9
Orientações *cis* (acima) e *trans* (em baixo) das ligações glicosídicas RNA. Figura da referência 20.

ção do tipo A, ou seja A-RNA, similar a A-DNA na figura 1.4. Diferente do B-DNA, que predomina para DNA, onde as bases são essencialmente perpendiculares ao eixo principal da hélice, no A-RNA há uma importante inclinação chamado de *tilt* [19]. No entanto, em geral o RNA na célula não é uma fita dupla bem formada como no DNA, e sim uma fita simples que dobra em sí mesma e assume conformações bem mais complicadas que serão discutidas nos capítulos 5 e 6. O pareamento canônico tipo Watson-Crick do RNA é muito similar ao de DNA, com citosina-guanina (CG) formando três ligações de hidrogênio e adenina-uracila (AU) com duas ligações como mostrado na figura 1.7. Mas mesmo quando o RNA forma uma fita dupla tipo A outros tipos pareamentos são muito comuns. Para entender esses pareamentos precisamos primeiro introduzir as faces de interação entre bases que são possíveis no RNA, ilustrados na figura 1.8, temos a face Watson-Crick já conhecida, a face tipo Hoogsteen e a face tipo açúcar (*sugar edge*) [20]. Além destas faces de interação, pode haver rotação das bases em relação às ligações glicosídicas como ilustrado na figura 1.9.

Enquanto o DNA é o código onde estão armazenadas as informações genéticas do indivíduo, o

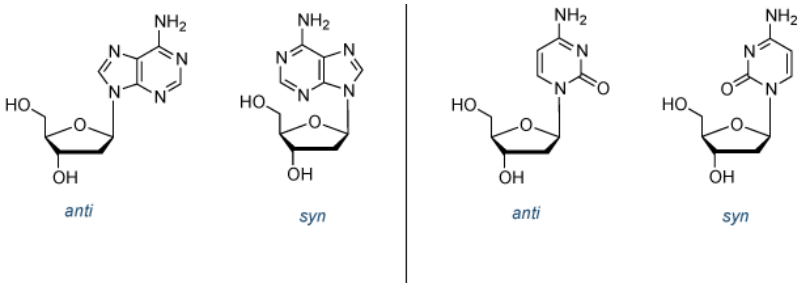


Figura 1.10
Conformações syn e anti para purinas (deoxiadenosina, parte esquerda) e pirimidinas (deoxicitosina, parte direita). Figura da referência 22.

RNA possui mais de uma função[2]. Ele transcreve o DNA, transporta aminoácidos, regula a quantidade de proteína sendo produzida, participa da formação de organelas (por exemplo, ribossomos), e diversas outras. No capítulo 5 revisaremos com mais detalhe as funções biológicas do RNA.

1.5 Conformações dos pareamentos

Os nucleotídeos ao se parearem, quer sejam eles RNA ou DNA, irão estar em uma configuração específica. Ela pode ser chamada *syn* ou *anti*, como discutido no trabalho de Kang et. al. [21] para o caso de DNA de telômeros. A figura 1.10 ilustra o que muda em cada configuração.

1.6 Conclusão

Nesta introdução mostramos os principais aspectos estruturais de DNA e RNA que serão usados nos próximos capítulos. Em particular mostramos as conformações estruturais complexas que RNA pode assumir quando comparado com o DNA.

Capítulo 2

Termodinâmica de experimentos de desnaturação

Ao longo desta tese faremos continuamente referências aos parâmetros termodinâmicos utilizados na descrição do DNA no contexto do modelo de próximos vizinhos (*nearest-neighbour model*, NN). Este modelo é muito usado na literatura para descrição de ácidos nucleicos e explicaremos seus pormenores através de um exemplo, na seção 2.1. Em seguida, na seção 2.2 apresentaremos a explicação do experimento que nos permite obter os parâmetros a serem usados em tal modelo, assim como os cálculos e princípios teóricos que nos permitem extrair tais parâmetros naquela situação.

2.1 Experimento de desnaturação de DNA

Por desnaturação (*melting*) entendemos o processo de separação das duas fitas que compõe a hélice dupla do DNA. Já o processo inverso, a formação da hélice a partir de duas fitas simples é chamado de hibridização. Em solução, as temperatura de desnaturação e de hibridização são aproximadamente iguais para baixas concentrações de DNA. A temperatura de desnaturação T_m de DNA em solução é determinada a partir de experimentos de absorção ótica ultra-violeta (UV) na faixa de 260 nm [23]. Em baixas temperaturas, o DNA da solução é formado predominantemente por fitas duplas (duplexos) e em altas temperaturas por fitas isoladas enoveladas (DNA desnaturado). A absorção ótica no UV aumenta na medida em que há mais fitas isoladas como ilustrado na figura figura 2.1, pois a parte das bases nitrogenadas que se ligam uma na outra absorvem luz nesta faixa.

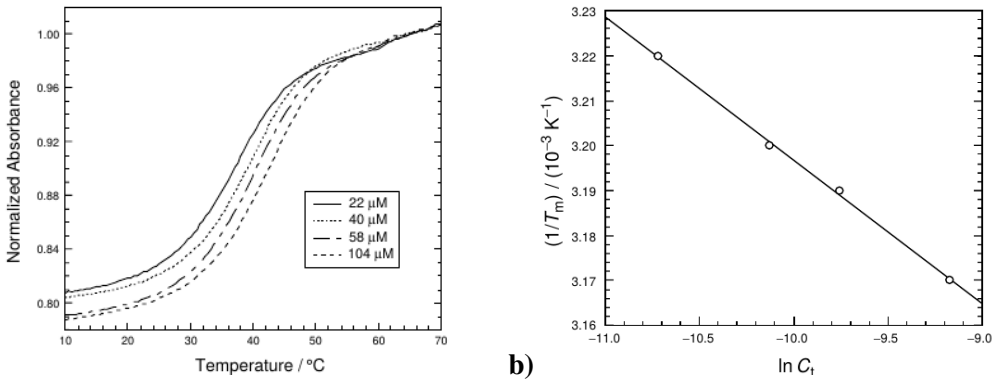


Figura 2.1
Variação da absorção UV (a) e da temperatura $T_m^{-1}s$ (b) para varias concentrações totais C_t. Figura da referência [23].

A temperatura de desnaturação T_m pode ser determinada diretamente por uma curva do tipo da figura 2.1 calculando a derivada e determinando o seu ponto de inflexão - este método é válido para a situação que respeita os critérios que descreveremos a seguir. O valor desta temperatura depende de diversos fatores, sendo o principal a concentração total C_t de DNA na solução. Tipicamente T_m^{-1} varia linearmente com $\ln C_t$, como visto na figura 2.1b. Quando esta relação se verifica é dito que o DNA tem uma desnaturação em dois estados (*two-state melting*). No entanto, nem sempre isto ocorre desta maneira, especialmente para sequências mais longas. Em seguida mostraremos os princípios da desnaturação de dois estados.

2.1.1 Desnaturação por dois estados

Vamos assumir que duas fitas de DNA complementares (digamos, fita X e fita Y) se apresentam de duas formas em uma solução aquosa: entrelaçadas formando uma dupla hélice (que representaremos como X·Y), ou como fitas separadas. Ou seja, um estado respresenta as fitas separadas X e Y, e o outro estado as fitas juntas X·Y. Podemos definir a seguinte reação reversível

$$X \cdot Y \rightleftharpoons X + Y \tag{2.1}$$

onde $[X]$ e $[Y]$ são as concentrações dos dois tipos de fitas e $[X \cdot Y]$ a concentração de duplexos [24]. Uma grandeza importante é a concentração total C_t *inicial* das fitas, que permanece constante durante todo o experimento, e que é dada por

$$C_t = [X]_0 + [Y]_0 \tag{2.2}$$

onde $[X]_0$ e $[Y]_0$ são as concentrações antes de ocorrer qualquer hibridização. Ocorrendo a reação da Eq. (2.1), a concentração total é dada por

$$C_t = [X]_t + [Y]_t + 2[X \cdot Y]_t \tag{2.3}$$

Na discussão que segue, é importante lembrar que C_t é sempre a concentração *inicial*.

Assumindo que só há dois estados para o DNA, fita separada ou duplexo, podemos associar a variação da energia livre de Gibbs ΔG em uma dada temperatura T com a constante de equilíbrio da reação K_{eq}

$$\Delta G = \Delta H - T\Delta S = -RT \ln K_{\text{eq}} \tag{2.4}$$

onde R é a constante dos gases ideais, ΔH a variação de entalpia e ΔS a variação de entropia [24]. A constante de equilíbrio químico desta reação de desnaturação é dado por

$$K_{\text{eq}} = \frac{[\text{produtos}]}{[\text{reagentes}]} = \frac{[X]\,.\,[Y]}{[X.Y]} \tag{2.5}$$

A temperatura de desnaturação T_m é definida como sendo aquela em que metade das fitas está desnaturada, ou seja

$$[X]_m = [Y]_m = [X \cdot Y]_m. \tag{2.6}$$

Podemos então escrever a concentração total C_t da Eq. (2.3) em função de uma das concentrações apenas, digamos $[X]_m$, de modo que $C_t = 4[X]_m$. Com isto a constante de equilíbrio se torna

$$K_{\text{eq}} = [X]_m = \frac{C_t}{4} \tag{2.7}$$

Finalmente, substituindo 2.7 em 2.4, a temperatura de desnaturação será expressa como

$$T_m = \frac{\Delta H}{\Delta S - R \ln \frac{C_t}{4}} \tag{2.8}$$

Auto-complementaridade: Para sequências auto-complementares a equação (2.8) pode mudar. Por exemplo a sequência 5′-ACGT-3′ complementa com 3′-TGCA-5′, ou seja é a própria sequência em reverso. Neste caso não temos dois tipos de fitas distintas X e Y, mas apenas fitas de um tipo somente, digamos X. Assim, a reação será $X \cdot X \rightleftharpoons 2X$ e a concentração inicial será $C_t = [X]_0$. Seguindo o mesmo tipo de desenvolvimento chegamos a

$$T_m = \frac{\Delta H}{\Delta S - R \ln C_t} \tag{2.9}$$

ou seja, similar a equação (2.8) a menos do fator $1/4$ na concentração.

Expressando a equação (2.8) pelo inverso da temperatura

$$T_m^{-1} = \frac{\Delta S}{\Delta H} - \frac{R}{\Delta H} \ln \frac{C_t}{4} \tag{2.10}$$

vemos que há uma relação linear entre T_m^{-1} e o logaritmo da concentração $\ln \frac{C_t}{4}$. É o que observamos na figura 2.1b que é chamado de gráfico de Van't Hoff (*Van't Hoff plot*) [24]. No entanto, não é grantido que um duplexo de DNA desnature desta maneira. No caso em que a relação $T_m^{-1} \times \ln C_t$ não é linear (*non-two state melting*), a função principal da equação (2.10) é determinar se a sequência de DNA em questão desnatura por um processo de dois estados. Neste caso pode-se obter os valores de ΔH e ΔS por regressão linear. O conhecimento dos valores de ΔH e ΔS de um conjunto grande de sequências diferentes de DNA permite prever a temperatura de desnaturação de sequências desconhecidas usando o modelo de próximos vizinhos que apresentaremos na seção seguinte.

2.2 Modelo de próximos vizinhos

A ideia do modelo de modelo de próximos vizinhos (*nearest-neighbour model*, NN) é atribuir valores de ΔH_α e ΔS_α para cada tipo de combinação α de próximos vizinhos. O valor total destas variações é obtido então pela simples soma para uma dada sequência i

$$\Delta H_i^{\text{tot}} = \sum_\alpha n_{i\alpha} \Delta H_\alpha + \sum_c n_i^c \Delta H^c \tag{2.11}$$

$$\Delta S_i^{\text{tot}} = \sum_\alpha n_{i\alpha} \Delta S_\alpha + \sum_c n_i^c \Delta S^c \tag{2.12}$$

onde $n_{i\alpha}$ é o número de ocorrências de próximos-vizinhos do tipo α. ΔH^c e ΔS^c são fatores de correção adicionais que não são do tipo próximos-vizinhos. Usando a equação (2.4) pode ser definido uma variação total de energia livre Gibbs em uma temperatura de referência, tipicamente 25°C ou 37°C,

$$\Delta G_i^{\text{tot}}(T) = \Delta H_i^{\text{tot}} - T \Delta S_i^{\text{tot}} \tag{2.13}$$

A energia livre de Gibbs à 37°C, $\Delta G_i^{\text{tot}}(37^\circ\text{C})$, é usada na determinação de estruturas secundárias como veremos no capítulo 6.

No caso de DNA canônico, há um total de 4^2=16 combinações α possíveis de próximos vizinhos, que podem ser reduzidos a 10 usando considerações de simetria. As simetrias no caso são avaliadas levando em consideração as direções 5′-3′ da fita dupla, veja a figura 1.5 da página 5. Os valores de ΔH_α e ΔS_α, para as 10 configurações α, podem então ser deduzidos dos valores totais usando métodos algébricos [25] ou de otimização [26].

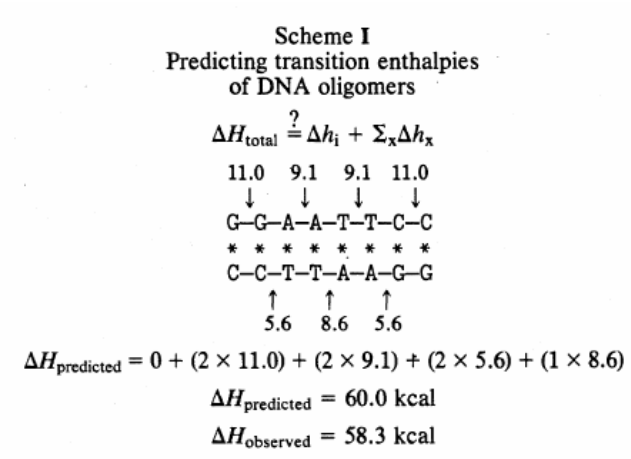


Figura 2.2
Ilustração de uma sequência de DNA $5'GGAATTCC3'$ e sua decomposição em próximos vizinhos. Figura da referência 27.

Para entender como o modelo funciona, tomamos como exemplo uma sequência qualquer de DNA, por exemplo 5′-GGAATTCC-3′, onde cada base está pareada com seu correspondente canônico, como mostrado na figura 2.2. Na figura vemos os pares de base, com os nucleotídeos da fita $5' \rightarrow 3'$, na parte de cima fazendo pares com os da fita $3' \rightarrow 5'$ na parte de baixo [27]. Notamos que esta sequência em particular é auto-complementar. Os pareamentos são representados com $*$, e os empilhamentos são representados com um traço simples. Note que cada empilhamento corresponde a uma adição na energia de formação do seu composto de DNA. Para entender o funcionamento do modelo de próximos vizinhos precisamos observar as equações presentes na figura 2.2. Note que a energia de formação do composto é descrito através do somatório das contribuições energéticas vindas de cada tipo de vizinho. A energia que caracteriza cada interação de vizinhança está indicada por setas que apontam para o espaço entre bases na figura 2.2 (ou seja, um par de bases GC vizinha a um par de bases GC contribui 11 kcal de energia para o sistema final). Com os valores previstos na figura 2.2 pode-se chegar à energia ΔH predita de 60 kcal, em comparação com os 58.3 kcal que foram observados experimentalmente. Para obter os valores de entropia procede-se de maneira idêntica, e de posse dos dois valores é possível aplicar a equação (2.8) e prever a temperatura de desnaturação T_m para quaisquer outras sequências. Embora a concordância seja muito boa, para fins práticos ainda é necessário intoduzir outros fatores de correção tais como fatores de inciação (como ΔS^{init}), fatores terminais (como ΔH^{term}) e correção de simetria (ΔS^{symm}) [25, 28]. A tabela 2.1 mostra os valores destes parâmetros calculados usando diversos métodos e usando diferentes conjuntos de dados experimentais para DNA [26]. De maneira inteiramente análoga, temos os parâmetros de próximos-vizinhos para RNA [29] mostrados na tabela 2.2. Aqui a principal diferença é a presença de um fator de correção para bases terminais AU que desestabiliza o duplexo de RNA.

Tabela 2.1
Conjuntos de ΔH_α e ΔS_α para DNA. Os parâmetros P-SL96 e P-SL98 são das referências 28 e 25 respectivamente. Os demais parâmetros são otimizações realizadas a partir de vários conjuntos de dados da referência 26.

Parâmetro	P-SL96	AOP-SL96	P-SL98	AOP-SL98	AOP-OW04	AOP-OW04DSC	AOP-CMB
$\langle\Delta H_{ApA}\rangle$	-8.4	-7.2(1)	-7.9	-8.4(1)	-8.6(2)	-8.2(4)	-7.20(4)
$\langle\Delta H_{ApC}\rangle$	-8.6	-10.4(1)	-8.4	-10.2(1)	-1.8(2)	-8.3(3)	-7.26(4)
$\langle\Delta H_{ApG}\rangle$	-6.1	-6.8(2)	-7.8	-5.3(1)	-7.0(3)	-6.4(3)	-6.46(5)
$\langle\Delta H_{ApT}\rangle$	-6.5	-9.5(2)	-7.2	-0.4(2)	-2.2(3)	-3.5(6)	-5.69(7)
$\langle\Delta H_{CpA}\rangle$	-7.4	-3.2(1)	-8.5	-12.9(2)	-15.0(2)	-8.5(3)	-8.58(6)
$\langle\Delta H_{CpC}\rangle$	-6.7	-7.49(8)	-8.0	-8.9(2)	-11.0(2)	-7.6(5)	-8.77(6)
$\langle\Delta H_{CpG}\rangle$	-10	-5.5(1)	-11	-9.3(1)	-14.4(4)	-12.1(6)	-9.70(6)
$\langle\Delta H_{GpA}\rangle$	-7.7	-2.77(9)	-8.2	-10.6(1)	-9.5(3)	-11.7(3)	-7.10(6)
$\langle\Delta H_{GpC}\rangle$	-11	-9.1(1)	-9.8	-14.5(2)	-6.8(4)	-12.5(3)	-9.59(8)
$\langle\Delta H_{TpA}\rangle$	-6.3	1.8(1)	-7.2	-6.1(2)	-13.8(3)	-12.9(5)	-4.08(7)
$\langle\Delta S_{ApA}\rangle$	-24	-20.3(4)	-22	-23.8(3)	-24.3(6)	-23(1)	-20.2(1)
$\langle\Delta S_{ApC}\rangle$	-23	-28.6(4)	-22	-28.3(5)	-3.2(7)	-22.2(8)	-19.3(1)
$\langle\Delta S_{ApG}\rangle$	-16	-17.9(5)	-21	-13.0(4)	-18.5(8)	-16.8(7)	-17.1(2)
$\langle\Delta S_{ApT}\rangle$	-19	-26.6(5)	-20	0.9(5)	-5.6(9)	-10(2)	-15.9(2)
$\langle\Delta S_{CpA}\rangle$	-19	-8.3(4)	-23	-36.6(6)	-42.0(7)	-23.0(8)	-23.2(2)
$\langle\Delta S_{CpC}\rangle$	-16	-19.7(3)	-20	-22.1(5)	-29.3(7)	-19(1)	-22.7(2)
$\langle\Delta S_{CpG}\rangle$	-26	-12.9(3)	-27	-22.9(4)	-39(1)	-31(2)	-24.9(2)
$\langle\Delta S_{GpA}\rangle$	-20	-6.9(3)	-22	-29.4(5)	-26(1)	-32.0(9)	-18.9(2)
$\langle\Delta S_{GpC}\rangle$	-28	-23.9(3)	-24	-38.9(7)	-16(1)	-32.4(8)	-24.4(2)
$\langle\Delta S_{TpA}\rangle$	-19	4.6(5)	-21	-17.7(7)	-40.3(8)	-37(2)	-11.3(2)
$\langle\Delta S^{\text{init}}_{AT}\rangle$	-9.0	-6.0(2)	—	—	—	—	—
$\langle\Delta S^{\text{init}}_{CG}\rangle$	-5.9	-0.9(1)	—	—	—	—	—
$\langle\Delta H^{\text{term}}_{5'\text{AT}}\rangle$	0.40	-0.40(3)	—	—	—	—	—
$\langle\Delta S^{\text{symm}}\rangle$	-1.4	0.67(6)	-1.4	-0.30(4)	—	—	-0.59(3)
$\langle\Delta H^{\text{term}}_{CG}\rangle$	—	—	0.10	0.7(1)	4.9(4)	10.6(7)	0.49(4)
$\langle\Delta S^{\text{term}}_{CG}\rangle$	—	—	-2.8	-1.5(3)	12(1)	29(2)	-0.7(1)
$\langle\Delta H^{\text{term}}_{AT}\rangle$	—	—	2.3	1.3(2)	6.5(3)	12.6(8)	2.53(7)
$\langle\Delta S^{\text{term}}_{AT}\rangle$	—	—	4.1	0.2(5)	17(1)	35(3)	5.7(2)

Tabela 2.2
Conjuntos de $\Delta G(37^\circ\text{C})_\alpha$, ΔH_α, ΔS_α para RNA. Tabela da referência 29.

parameters	$\Delta G°_{37}$ (kcal/mol)	$\Delta H°$ (kcal/mol)	$\Delta S°$ [b] (eu)
5′AA3′ 3′UU5′	−0.93 (0.03)	−6.82 (0.79)	−19.0 (2.5)
5′AU3′ 3′UA5′	−1.10 (0.08)	−9.38 (1.68)	−26.7 (5.2)
5′UA3′ 3′AU5′	−1.33 (0.09)	−7.69 (2.02)	−20.5 (6.3)
5′CU3′ 3′GA5′	−2.08 (0.06)	−10.48 (1.24)	−27.1 (3.8)
5′CA3′ 3′GU5′	−2.11 (0.07)	−10.44 (1.28)	−26.9 (3.9)
5′GU3′ 3′CA5′	−2.24 (0.06)	−11.40 (1.23)	−29.5 (3.9)
5′GA3′ 3′CU5′	−2.35 (0.06)	−12.44 (1.20)	−32.5 (3.7)
5′CG3′ 3′GC5′	−2.36 (0.09)	−10.64 (1.65)	−26.7 (5.0)
5′GG3′ 3′CC5′	−3.26 (0.07)	−13.39 (1.24)	−32.7 (3.8)
5′GC3′ 3′CG5′	−3.42 (0.08)	−14.88 (1.58)	−36.9 (4.9)
initiation[c]	4.09 (0.22)	3.61 (4.12)	−1.5 (12.7)
per terminal AU[d]	0.45[d] (0.04)	3.72[d] (0.83)	10.5[d] (2.6)
symmetry correction (self-complementary)	0.43	0	−1.4
symmetry correction (non-self-complementary)	0	0	0

Capítulo 3

Modelo Peyrard-Bishop

3.1 Introdução

O modelo Peyrard-Bishop é um modelo físico para análise estrutural das propriedades termodinâmicas de cadeias de DNA e RNA. Foi apresentado pela primeira vez em 1989 no artigo *Statistical mechanics of a nonlinear model for DNA denaturation* de autoria de Peyrard e Bishop [30]. Alguns modelos e métodos teóricos buscam obter as propriedades de uma dada sequência de DNA, e suas propriedades, através da descrição completa da mesma — incluindo as informações de cada átomo participante da fita (sua massa, carga eletrônica, posição, etc). Estes são ditos modelos microscópicos para descrição do DNA. Exemplos de tais modelos são o tratamento por teoria do funcional de densidade (*density functional theory*, DFT) [31] e por dinâmica molecular (*molecular dynamics*, MD) [32]. Outros modelos buscam descrever as propriedades da fita como um todo, ou em grandes blocos: um exemplo disso é a técnica de *worm-like-chain* [33]. Tais modelos são denominados macroscópicos. O PB, por sua vez, é um modelo intermediário. Não apresenta uma descrição atomística minuciosa, nem uma descrição da fita como um todo. Ele descreve o DNA usando como unidades as bases nitrogenadas, formadas por coleções de dezenas de átomos, e suas propriedades físicas. Por este motivo, ele é dito um modelo mesoscópico.

O modelo tem uma longa história de aplicação, como por exemplo o DNA *breathing* [34–36], predição de sítios de transcrição [37] e análise fractal de desnaturação [38], efeitos de forças externas [39], entre outros. Por sua versatilidade pode ser usado por vários métodos teóricos tais como dinâmica de Langevin [40], técnicas de integrais de caminho [41, 42], cálculo de coeficientes de Lyapunov [43] e análise por *wavelets* [44]. As adições feitas ao longo do tempo, que o permitem

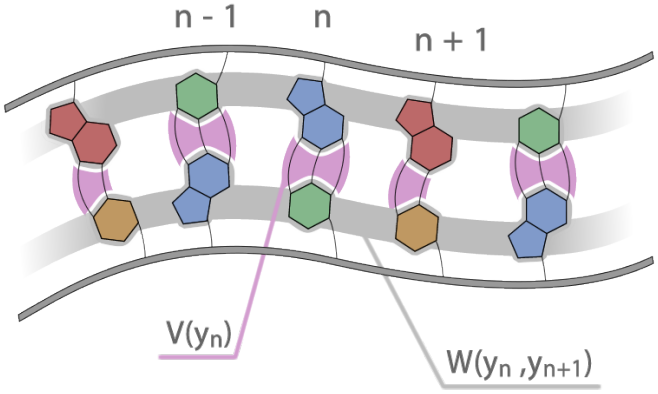


Figura 3.1
Representação esquemática do modelo Peyrard-Bishop.

analisar mais parâmetros estruturais como a torção tridimensional da fita [45, 46], *loops* e *hairpins* de pares de base [47], interação com solvente [48, 49], dentre outros [35], o torna uma ferramenta versátil para estudo deste sistema físico.

Nas seções seguintes vamos apresentar a base teórica do modelo, passando pelo cálculo de temperaturas de desnaturação e terminando com o procedimento de parametrização que será a base do capítulo 4.

3.2 O modelo Peyrard-Bishop para DNA homogêneo

A ideia básica do modelo é representar as ligações de hidrogênio e as interações por potenciais simples, que permitam montar uma Hamiltoniana com a qual seja possível calcular a função partição clássica. Aqui apresentaremos o desenvolvimento proposto por Peyrard e Bishop [30] onde se assume um DNA formado por duas fitas *planas* e homogêneo, ou seja formado somente um tipo de par de base. Assim, a Hamiltoniana 2D é escrita em termos de dois deslocamentos u e v para cada lado do DNA [30]

$$H = \sum_n \frac{1}{2}m\left[\dot{u}_n^2 + \dot{v}_n^2\right] + \frac{1}{2}k\left(u_n - u_{n-1}\right)^2 + \frac{1}{2}k\left(v_n - v_{n-1}\right)^2 + D\left[e^{-(u_n - v_n)/\lambda} - 1\right]^2 \quad (3.1)$$

onde m é a massa média do nucleotídeo. A interação de empilhamento é descrito por um potencial harmônico com constante elástica k. As ligações de hidrogênio são descritas por um potencial tipo Morse [50] onde D é a profundidade do potencial e o parâmetro λ_a está relacionado com a sua largura. Como se trata de um DNA homogêneo, os valores de k e D não mudam ao longo da fita.

Para simplificar a equação (3.1) se introduz a seguinte transformação de coordenadas

$$x_n = \frac{u_n + v_n}{\sqrt{2}}, \quad y_n = \frac{u_n - v_n}{\sqrt{2}} \tag{3.2}$$

o que resulta em

$$H = \sum_n \frac{1}{2} m \left[\dot{x}_n^2 + \dot{y}_n^2\right] + \frac{1}{2} k \left(x_n - x_{n-1}\right)^2 + \frac{1}{2} k \left(y_n - y_{n-1}\right)^2 + D \left[e^{-\sqrt{2} y_n / \lambda_n} - 1\right]^2, \tag{3.3}$$

note que o termo de potencial de Morse agora depende somente de y. Vamos agora expressar a hamiltoniana da equação (3.3) em termos de potenciais genéricos V e W

$$H = \sum_n \left[\frac{1}{2m} p_n^2 + \frac{1}{2m} q_n^2 + W(x_n, x_{n-1}) + W(y_n, y_{n-1}) + V(y_n)\right] \tag{3.4}$$

onde $p = m\dot{x}$ e $q = m\dot{y}$, e

$$W(y_n, y_{n-1}) = \frac{k}{2}(y_n - y_{n-1})^2 \tag{3.5}$$

e

$$V(y_n) = D\left(e^{-y_n/\lambda_n} - 1\right)^2 \tag{3.6}$$

onde a amplitude D_n representa quão fortes são as pontes de hidrogênio, e λ_n ajusta a largura do poço de potencial. A escolha pelo potencial de Morse é justificada pelo fato de ele apresentar o comportamento qualitativo esperado para a interação entre as bases que formam a ponte de hidrogênio. Especificamente, ele traz (a) repulsão para $y < 0$, caso em que as bases ficam muito próximas, (b) um mínimo de potencial em $y = 0$, que representa a distância em que as bases ficam em equilíbrio, e (c) a diminuição gradual para a interação entre as bases para valores cada vez maiores da distância $y > 0$ entre elas, até o limite em que elas não mais interagem.

Nós vamos usar a Hamiltoniana da equação (3.4) para calcular a função partição

$$Z = \int \prod_n e^{-\beta H} dp_n dq_n dx_n dy_n = Z_p Z_q Z_x Z_y \tag{3.7}$$

onde $\beta = 1/(k_B T)$, sendo k_B a constante de Boltzman e T a temperatura do sistema. As integrais em p_n, q_n e x_n podem ser calculadas

$$Z_p = Z_q = (2\pi m k_B T)^{N/2} \tag{3.8}$$

e

$$Z_x = \left(\frac{2\pi}{\beta k}\right)^{N/2} \tag{3.9}$$

onde N é o número de pares de bases da sequência de DNA. Ou seja, Z_p, Z_q e Z_x são constantes, e independentes dos potenciais.

Resta o termo em y

$$Z_y = \int \prod_n e^{-\beta H_y} dy_n \tag{3.10}$$

com

$$H_y = \sum_n \left[W(y_n, y_{n-1}) + V(y_n) \right] \tag{3.11}$$

ou seja, na prática nossa Hamiltoniana depende apenas da variável y razão pela qual muitas vezes o modelo PB também ser considerado um modelo unidimensional (1D) apesar da Hamiltoniana da qual partimos, equação (3.3), ser claramente bidimensional. A forma da equação (3.11) agora deixa mais claro a representação esquemática da figura 3.1.

3.2.1 O método da matriz de transferência

Prosseguimos no cálculo de Z_y, escrevendo explícitamente as integrais

$$Z_y = \int dy_1 \int dy_2 \ldots \int dy_N \prod_n e^{-\beta W(y_n, y_{n-1})} e^{-\beta V(y_n)} \tag{3.12}$$

Ou seja, nos resta resolver uma integral N dimensional de um produtório de exponenciais. Este tipo de integral pode ser tratado com a técnica de matriz de transferência [51]. Primeiro fazemos um rearranjo dos termos

$$\begin{aligned} Z_y = \int dy_1 \int dy_2 \ldots \int dy_N \qquad & e^{-\beta W(y_1, y_N)} e^{-\frac{\beta}{2}[V(y_1)+V(y_N)]} \\ & \times e^{-\beta W(y_2, y_1)} e^{-\frac{\beta}{2}[V(y_2)+V(y_1)]} \\ & \vdots \\ & \times e^{-\beta W(y_N, y_{N-1})} e^{-\frac{\beta}{2}[V(y_N)+V(y_{N-1})]} \end{aligned} \tag{3.13}$$

com isto simetrizamos os potenciais

$$U(y_i, y_j) = W(y_i, y_j) + \frac{1}{2}\left[V(y_i) + V(y_j) \right] \tag{3.14}$$

de maneira que $U(y_i, y_j) = U(y_j, y_i)$. Agora introduzimos a função *kernel* K

$$K(y_i, y_j) = e^{-\beta U(y_i, y_j)} \tag{3.15}$$

que também satisfaz $K(y_i, y_j) = K(y_j, y_i)$. A integral apresentada na equação (3.13) pode agora ser escrita em termos desta função *kernel*

$$Z_y = \int dy_1 \int dy_2 \ldots \int dy_N K(y_1, y_N) K(y_2, y_1) K(y_3, y_2) \ldots K(y_N, y_{N-1}) \tag{3.16}$$

Introduzimos a equação integral

$$\int K(y_i, y_j)\phi(y_i)dy_i = \lambda\phi(y_j) \tag{3.17}$$

onde se assume

$$||K(y_i, y_j)|| = \left(\int\int [K(y_i, y_j)]^2 dy_i dy_j\right)^{1/2} < \infty \tag{3.18}$$

com $K(y_i, y_j) > 0$, neste caso a integral tem um kernel Hilbert-Schmidt positivo, e tem um conjunto de autovalores positivos e autovetores ortonormais. Ordenando os autovalores em ordem descrescente $\lambda_1, \lambda_2, \ldots$ e seus respectivos autovetores $\phi_1(x), \phi_2(x), \ldots$

$$\langle n|m\rangle = \int \phi_n(x)\phi_m(x)dx = \delta_{n,m} \tag{3.19}$$

e

$$\sum_{n=1}^{\infty} \phi_n(x)\phi_n(y) = \delta(x - y) \tag{3.20}$$

a função $K(y_i, y_j)$ pode ser expandida

$$K(y_i, y_j) = \sum_n \lambda_n\phi_n(y_i)\phi_n(y_j) \tag{3.21}$$

Para exemplificar o procedimento de integração, fazemos a integração em y_2

$$\begin{aligned}\int K(y_1, y_2)K(y_2, y_3)dy_2 &= \int \sum_n \lambda_n\phi_n(y_1)\phi_n(y_2) \sum_m \lambda_m\phi_m(y_2)\phi_m(y_3)dy_2 \\ &= \sum_n\sum_m \lambda_n\lambda_m\phi_n(y_1)\phi_m(y_3)\delta_{n,m} \\ &= \sum_n \lambda_n^2\phi_n(y_1)\phi_n(y_3)\end{aligned} \tag{3.22}$$

prosseguimos desta maneira até chegar em

$$Z_y = \int \sum_n \lambda_n^N\phi_n(y_1)\phi_n(y_1)dy_1 = \sum_{n=1}^{\infty} \lambda_n^N\delta_{n,n} = \sum_{n=1}^{\infty} \lambda_n^N \tag{3.23}$$

ou seja, a equação (3.13) resume-se simplesmente em encontrar os autovalores λ_n [30].

3.2.2 Estiramento médio entre as bases

De maneira completamente análoga à equação (3.23), é possível calcular o estiramento médio das bases, que é a grandeza de interesse

$$\langle y_m\rangle = \frac{1}{Z}\int \prod_n y_m e^{-\beta H} dy_n \tag{3.24}$$

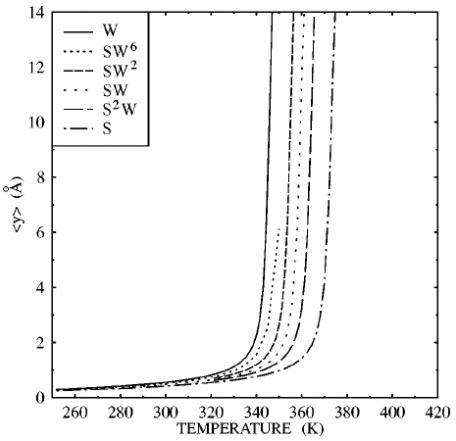


Figura 3.2
Valores de $\langle y \rangle$, relacionados ao estiramento das ligações de hidrogênio, em função da temperatura T para diferentes composições de sequências. W (*weak*, ou AT) e S (*strong*, ou CG) podem ser alternados na mesma sequência Figura da referência 52.

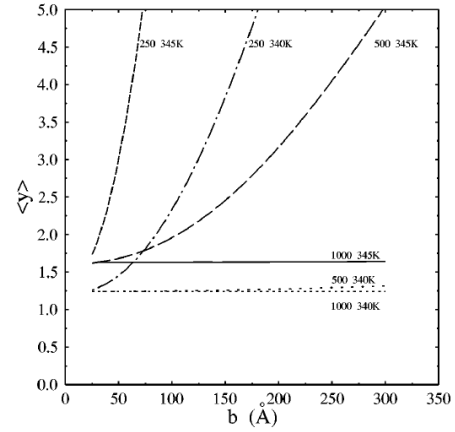


Figura 3.3
Valores de $\langle y \rangle$, em função do limite superior de integração b. Figura da referência 52.

e que se reduz a

$$\langle y \rangle = \frac{1}{Z_y} \sum_{n=1}^{\infty} \lambda_n^N \langle n|y|n \rangle \tag{3.25}$$

onde o índice m foi removido uma vez que todas as bases são iguais. A figura 3.2 mostra o resultado do cálculo numérico apresentando por Peyrard e Bishop [30] do valores de estiramento médio $\langle y \rangle$ em função da temperatura.

3.2.3 O problema da divergência

Zhang et al. [52] mostraram em 1997 um problema associado à forma da Hamiltoniana da equação (3.11) que resulta em uma divergência no cálculo da integral da equação (3.17). Estes autores notaram que ao variar o limite de integração da integral da equação (3.17) o valor de $\langle y \rangle$ poderia aumentar continuamente, algo que não deveria acontecer, veja a figura 3.3. A razão para esta divergência reside na forma do potencial de empilhamento, equação (3.5), no caso particular onde $y_n = y_{n-1}$ o potencial é zero, $W = 0$, e isto leva à divergência da própria integração numérica da equação (3.13). Este problema aparece de maneira particularmente grave no caso do empilhamento anarmônico proposto por Dauxois et al. [53, 54]. Uma maneira de contornar este problema

foi proposto por nosso grupo [55], introduzindo um ângulo θ entre y_n e y_{n-1}, de maneira análoga às Hamiltonianas 3D [45, 46],

$$W(y_n, y_{n-1}) = \frac{k_n}{2}(y_n^2 - 2y_n y_{n-1}\cos\theta + y_{n-1}^2) \tag{3.26}$$

onde mantemos o ângulo fixo num valor pequeno $\theta = 0.01$ rad. Esta maneira de evitar a divergência não é única (i.e. veja Theodorakopoulos et al. [56]).

3.3 O modelo Peyrard-Bishop para DNA heterogêneo

Raramente o DNA é composto somente por um tipo de base, isto é, composto somente por pares CG ou somente AT. Então para que possa ter aplicação prática o método mostrado na seção anterior precisa ser estendido para DNA com qualquer tipo de composição. A função *kernel* K da equação (3.15) agora passa a ser da seguinte forma

$$K^{(i,i+1)}(y_i, y_{i+1}) = \exp\left(-\beta\left\{W_i(y_i, y_{i+1}) + \frac{1}{2}\left[V_i(y_i) + V_{i+1}(y_{i+1})\right]\right\}\right) \tag{3.27}$$

observe que agora os índices i e $i+1$ também afetam V e W, ou seja, não são mais as funções com os mesmos parâmetros k e D, mas com parâmetros que podem mudar de acordo com o sítio i, λ_i, k_i e D_i. Com isto muitas das simplificações da seção anterior não se aplicam mais. No entanto, Zhang et al. [52] propuseram um método interessante onde a função *kernel* K da equação (3.15) da sequência heterogênea é expandida da seguinte maneira

$$K^{(i,i+1)}(x, y) = \sum_{n,m=1}^{\infty} C_{nm}^{(i,i+1)}\phi_n(x)\phi_m(y) \tag{3.28}$$

com

$$C_{nm}^{(i,i+1)} = \int\int K^{(i,i+1)}(x, y)\phi_n(x)\phi_m(y)dxdy \tag{3.29}$$

onde ϕ são as autofunções da sequência homogênea. Desta maneira a função partição passa a ser produto de matrizes

$$Z_y = \mathrm{Tr}\left(\mathbf{C}^{(1,2)}\mathbf{C}^{(2,3)}\ldots\mathbf{C}^{(N,1)}\right) \tag{3.30}$$

onde as matrizes $\mathbf{C}$ são

$$\mathbf{C}^{(i,i+1)} = [C_{nm}^{(i,i+1)}] \tag{3.31}$$

De maneira inteiramente análoga, o estiramento médio se torna

$$\langle y_i\rangle = \frac{1}{Z_y}\mathrm{Tr}\left(\mathsf{C}^{(1,2)}\mathsf{C}^{(2,3)}\ldots\mathsf{C}^{(i-1,i)}\mathsf{Y}\mathsf{C}^{(i,i+1)}\ldots\mathsf{C}^{(N,1)}\right) \tag{3.32}$$

onde os elementos da matrix Y são

$$Y_{nm} = \int \phi_n(y) y \phi_m(y) dy \tag{3.33}$$

Note que a posição de Y no produto varia conforme o índice i.

Como estas matrizes são tipo próximos-vizinhos (*nearest-neighbours*) já que dependem de um sítio i e o seu vizinho $i+1$. Como estas matrizes dependem apenas dos potenciais V e W, temos apenas 16 tipos destas matrizes sendo que considerações de simetriza ainda reduzem este número a 10 para DNA, assim como ocorreu com os parâmetros do modelo NN na seção 2.2, página 12. Temos então as 10 matrizes

$$\begin{aligned}
&\mathbf{C}^{AT,TA} && \mathbf{C}^{TA,AT} \\
&\mathbf{C}^{CG,GC} && \mathbf{C}^{GC,CG} \\
&\mathbf{C}^{AT,AT} = \left(\mathbf{C}^{TA,TA}\right)^T && \mathbf{C}^{CG,AT} = \left(\mathbf{C}^{TA,GC}\right)^T \\
&\mathbf{C}^{AT,CG} = \left(\mathbf{C}^{GC,TA}\right)^T && \mathbf{C}^{AT,GC} = \left(\mathbf{C}^{CG,TA}\right)^T \\
&\mathbf{C}^{GC,AT} = \left(\mathbf{C}^{TA,CG}\right)^T && \mathbf{C}^{CG,CG} = \left(\mathbf{C}^{GC,GC}\right)^T
\end{aligned} \tag{3.34}$$

onde T denota a matriz transposta. Portanto, para qualquer sequência de DNA, não importando o seu comprimento N, as matrizes da equação (3.34) precisam ser calculados apenas uma única vez. Para calcular o estiramento médio de qualquer outra sequência basta reordenar o produto da equação (3.32).

3.3.1 Método da equivalência térmica

Mesmo sendo simples, as curvas de desnaturação $\langle y \rangle$ como da figura 3.3 ainda são bastante demoradas para calcular e os valores de temperatura de desnaturação não são muito satisfatórios. O nosso grupo desenvolveu um método semi-empírico que utiliza as matrizes da equação (3.31) para determinar um índice que é correlacionado a temperaturas medidas experimentalmente [48, 57].

As matrizes C são determinadas a partir da expansão de uma sequência homogênea da equação (3.29). Vamos supor que usamos a sequência homogênea composta somente por bases CG para calcular as autofunções, neste caso a matriz $\mathbf{C}^{CG,GC}$ será uma matriz diagonal com os elementos formados pelos autovalores λ da equação (3.23). Vamos explorar este fato separando todas as matrizes C da equação (3.31) em partes diagonal Λ e não-diagonal Δ

$$\mathsf{C}^{(i,i+1)} = \Lambda + \Delta^{(i,i+1)} \tag{3.35}$$

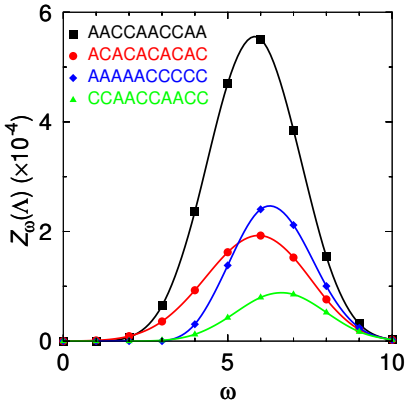


Figura 3.4
Componentes $Z_\omega(\Lambda)$, equação (3.38), para vários tipos de sequência. Note como o máximo da curva aparece para ω depende da ordem das bases. Figura da referência 57.

Se tivermos usado uma sequência homogênea CG teremos $\Delta^{CG,GC} = 0$ e

$$\mathbf{C}^{CG,GC} = \Lambda \tag{3.36}$$

e o produto da equação (3.30) toma a seguinte forma

$$Z_y = \mathrm{Tr}\left[(\Lambda + \Delta^{(1,2)})(\Lambda + \Delta^{(2,3)}) \ldots (\Lambda + \Delta^{(N,1)})\right] \tag{3.37}$$

que pode ser decomposto em $N+1$ componentes como função da ordem da matriz Λ

$$Z_y = \sum_{\omega=0}^{N} Z_\omega(\Lambda) = \sum_{\omega=0}^{N} \mathrm{Tr}[O(\Lambda^\omega)] \tag{3.38}$$

onde $O(\Lambda^\omega)$ são todos os termos da expansão contendo ω multiplicações da matriz Λ.

Na figura 3.4 mostramos um exemplo desta expansão, note como a curva muda, em particular o valor ω para o qual a curva é máximo depende da estrutura da sequência. Se usarmos este valor, $\omega_{\max}$ observamos que não há uma dependência forte com a temperatura como se vê na figura 3.5. Ou seja, $\omega_{\max}$ é um valor que reflete a *estrutura* da sequência de DNA e essencialmente independente de temperatura. Isto nos permite usar $\omega_{\max}$ como um indexador e correlacioná-lo com temperaturas de desnaturação. A figura 3.6 mostra que $\sqrt{\omega_{\max}}$ tem uma relação linear com temperaturas de desnaturação de Owczarzy et al. [58], isto nos permite então realizar a regressão linear para cada grupo de comprimento N

$$T_p = a_0(N) + a_1(N)\tau, \tag{3.39}$$

onde $\tau = \sqrt{\omega_{\max}}$. De posse dos coeficientes $a_{0,1}$ podemos usar a equação (3.39) para prever a temperatura de desnaturação T_p para quaisquer outras sequências de DNA.

3.3.2 Procedimento de cálculo de desnaturação

Podemos resumir o procedimento de cálculo de desnaturação da seguinte maneira:

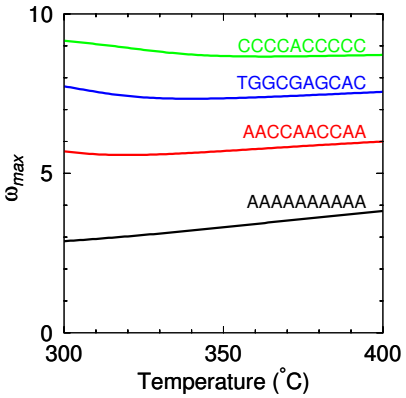


Figura 3.5
Valor de $\omega_{\max}$ em função de temperatura para várias sequências de DNA. Figura da referência 57.

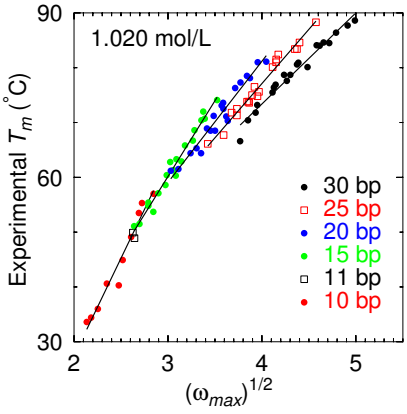


Figura 3.6
Relação entre $\omega_{\max}$ e temperatura de desnaturação experimentais. As sequências são agrupadas por tamanho (10–30 pares de base) e a regressão linear é feito para cada grupo separadamente. Figura da referência 57.

1. estabelecemos os parâmetros da Hamiltoniana H_y, equação (3.11),

2. calculamos as matrizes da equação (3.34)

3. para conjunto de sequências, determinamos os valores de $\omega_{\max}$ pela expansão da equação (3.38)

4. calculamos os coeficientes $a_{0,1}$ da regressão linear da equação (3.39)

Para qualquer outra sequência de DNA, basta calcular o seu respectivo $\omega_{\max}$ e aplicar a equação equação (3.39), já que não há necessidade de recalcular as matrizes do passo 2. Com este procedimento, realizar predições de temperatura de desnaturação é computacionalmente quase tão eficiente quanto usar o método de próximos vizinhos que descrevemos na seção 2.2, página 12, com a diferença que podemos usar parâmetros associados à ligação de hidrogênio e interação de empilhamento. Por outro lado, em geral estes parâmetros não são em geral bem conhecidos, especialmente para outros tipos de ácidos nucleicos, porém podemos usar do mesmo procedimento descrito aqui para melhorar o valor dos parâmetros como vamos mostrar na próxima seção.

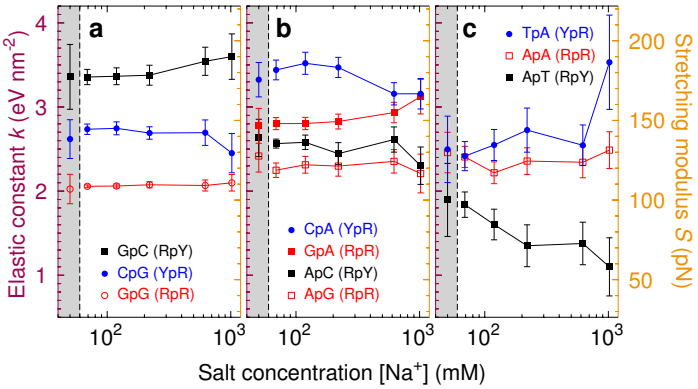


Figura 3.7
Constantes elásticas k otimizadas em função de concentração salina. Figura da referência 59.

3.4 Parametrização do modelo PB

Vamos usar o procedimento descrito na seção 3.3.2 para obter novos parâmetros para oligonucleotideos. A ideia é simplesmente variar o conjunto de parâmetros da Hamiltoniana H_y, equação (3.11), principalmente D e k, e procurar minimizar a diferença entre as temperaturas

$$\chi^2 = \sum_{i=1}^{N} \left(T_i' - T_i\right)^2 , \tag{3.40}$$

onde T_i' se refere à temperatura predita pelo modelo para a fita i e T_i se refere à temperatura experimental [59]. Desta maneira o problema de obter parâmetros passa a ser um problema de minimização. No caso de DNA canônico, teremos pelo menos dois parâmetros D e 10 parâmetros k, ou seja o espaço de busca é no mínimo de 12 parâmetros. Assim, a eficiência do procedimento da seção 3.3.2, que leva poucos segundos de processamento, é crucial para conseguir realizar esta minimização.

No entanto, o mais interessante do método não é de melhorar a predição das temperaturas de desnaturação. O importante é a constatação de que os novos parâmetros D e k, que agora são oriundos de dados experimentais de desnaturação por causa da equação (3.39) e portanto tem o potencial de revelar informações intramoleculares do DNA. Na figura 3.7 mostramos um exemplo onde os parâmetros k, que estão associados à elasticidade do DNA foram otimizados usando o procedimento da seção 3.3.2 e mostra que os próximos-vizinhos do tipo AT-AT são substancialmente mais flexíveis que CG-CG [59]. O método também foi aplicado com sucesso para RNA onde se mostrou que as ligações de hidrogênio da AU são mais fortes que de AT em DNA [60], corroborando resultados experimentais de NMR [61].

3.5 Conclusão

Neste capítulo apresentamos a base teórica do modelo PB e o procedimento de parametrização. Como vimos, a parametrização é muito mais do que a simples obtenção de valores bons para o modelo, ela serve como método para extrair informações intramoleculares. Mas o método como tal ainda está na sua infância e há muitas perguntas a serem respondidas e ainda há muita validação para ser feita. Em particular, fica a pergunta se o método de parametrização pode ser aplicado a oligonucleotídeos diferentes dos canônicos AT e CG em DNA. No capítulo 4 vamos tratar exatamente de um caso importante que é a inosina em DNA. Considerada uma "base universal", com um número de ligações de hidrogênio uniforme para qualquer configuração de pareamento, a inosina é um importante passo na validação do método.

Capítulo 4

Inosina em DNA pelo modelo Peyrard-Bishop (PB)

Neste capítulo apresentarei o desenvolvimento do primeiro trabalho que realizamos após minha entrada no Grupo de Biofísica Computacional. Utilizamos o modelo Peyrard-Bishop (PB), discutido no capítulo 3, para investigar as propriedades estruturais da base nitrogenada conhecida como *deoxyinosine*, mais comumente denominada apenas inosina. Nós obtivemos os parâmetros do potencial de Morse e do empilhamento para a interação da inosina com as diversas bases canônicas do DNA (i.e. A,C,G,T). Este trabalho culminou no artigo que se encontra no apêndice D [62].

4.1 A estrutura da Inosina

Inosina é uma base nitrogenada que ocorre naturalmente em RNA de transferência (*transfer RNA*, tRNA). Ela é formada pela hipoxantina quando ligada à ribose, sendo que a hipoxantina possui uma forma muito similar à guanina como pode ser visto na figura 4.1 tendo por única diferença a ausência do grupo amino. Quando a hipoxantina é ligada à desoxiribose ela forma a desoxiinosina, ou seja inosina em DNA. Uma caraterística especial da inosina é que ela apresenta pareamento estável tanto com citosina como também com adenina e uracila. De maneira análoga, a desoxiinosina apresenta pareamento estável com citosina, guanina e timina, como mostrado na figura 4.2 [63, 64].

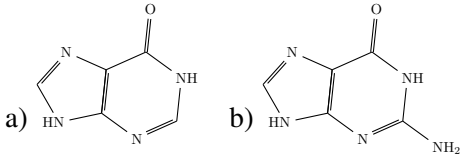


Figura 4.1
Estrutura da (a) hipoxantina (base da inosina) e da (b) guanina.

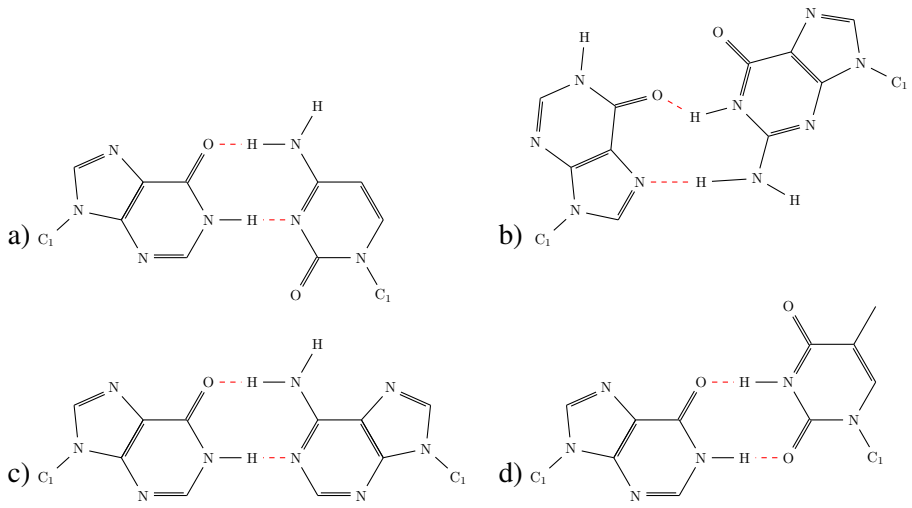


Figura 4.2
Estrutura de (a) inosina(anti)-citosina(anti), (b) inosina(syn)-guanina(anti), (c) inosina(anti)-guanosina(anti) e (d) inosina(anti)-timina(anti) em DNA.

Por simplicidade, nos vamos nos referir no restante deste capítulo à hipoxantina, inosina e desoxinosina pelo termo geral "inosina", com sigla I, como é comum de se encontrar na literatura, faremos a distinção apenas quando necessário.

A condição de pareamento generalizado fez com que a inosina fosse uma forte candidata à base “universal”, de forte interesse para a produção de *primers* para reação em cadeia da polimerase (*polymerase chain reaction*, PCR). Um histórico deste tipo de aplicação se inicia com a discussão de seu uso como sonda (*probe*) em 1985 [65, 66], segue com aplicações em PCR [67, 68], e, mais recentemente, tem sido usada em clonagem direcional [69], e sequenciamentos por ligações, onde ela atua com o propósito de aumento dos tamanhos de sequências que podem ser lidas por esta técnica [70, 71].

Os primeiros estudos a respeito da "universalidade"da complementaridade da inosina foram feitos por Kawase et al. [72], e Case-Green e Southern [73]. No entanto, o trabalho mais completo, e nossa principal fonte de dados experimentais para a modelagem que fizemos é proveniente de medidas de absorção de UV feitas por Watkins e SantaLucia [74]. A partir de medidas de energia livre do modelo de próximos vizinhos eles estabeleceram uma hierarquia de estabilidade que é usada em diversas aplicações práticas [69, 75]. Além das medidas de parâmetros de energia livre, também existem medidas de NMR [76, 77], e medidas de raios X [78], que fornecem informações sobre a estrutura dos compostos que contém inosina, e informações sobre as ligações químicas presentes nas mesmas. Todas estas medidas e estudos sugerem uma configuração de duas pontes de hidrogênio

para todos pareamentos da inosina mostrado na figura 4.2.

Um tratamento teórico completo para a inosina ainda não foi feito, mas, em contrapartida, existem estudos para a hipoxantina, que é a base nucleica que formará o nucleotídeo de inosina. A hipoxantina foi estudada com técnicas de DFT [79, 80]. Note que existe uma complicação em associar os resultado obtidos em tais estudos com os pareamento de inosina propriamente ditos, pois em tais cálculos não existe uma estrutura de cadeias de fosfato. Um estudo que inclui a cadeia de fosfato foi feito com dinâmica molecular por Krepl et al. [32] para uma única sequência de DNA contendo o pareamento IC. Neste estudo ele conclui que a estabilidade reduzida da sequência não se deve unicamente ao número reduzido de ligações de hidrogênio entre inosina e citosina (2 ligações de hidrogênio), em oposição a citosina com guanina (3 ligações de hidrogênio).

4.2 Motivação e objetivo

Como discutimos na seção 3.5, página 26, há um interesse em aplicar o modelo PB a outros tipos de oligonucleotídeos. Considerando o interesse tecnológico da inosina, a existência de dados experimentais de desnaturação [74], assim como a condição peculiar de formação uniforme de duas ligações de hidrogênio em qualquer configuração de pareamento, como mostrado na figura 4.2, esta base nos pareceu um objeto ideal de estudo. Assim, aqui nós vamos mostrar como adaptamos os dados de desnaturação e as estratégias de otimização adotadas. Em seguida, discutimos os resultados e fazemos uma correlação com o que se conhece da estrutura da inosina quando pareada com outras bases.

4.3 Dados de desnaturação

Para nossa modelagem PB iremos utilizar como dados experimentais as temperaturas de desnaturação de sequências contendo pareamentos com inosina, extraídos do trabalho de Watkins e SantaLucia [74]. Ao total temos 84 sequências originais deste trabalho específico, em conjunto com 13 sequências extraídas de outras fontes [66, 81]. Estas sequências foram escolhidas de tal maneira a cobrir todas interações com as vizinhanças possíveis para um dado pareamento de inosina. O erro experimental reportado para as temperaturas de desnaturação de tal trabalho é estimado em $0.3^{\circ}\mathrm{C}$ para as 84 sequências de Watkins e SantaLucia [74], todas elas com comprimentos entre 9 ou 12 pares de base. Não foram incluídas as sequências com par de inosina vizinho a outro par de inosina

(chamados de sequências *tandem*), pois o conjunto de sequências era pequeno demais para termos uma minimização de qualidade. Algumas sequências adicionais estavam presentes no trabalho de Watkins e SantaLucia [74], que o mesmo classificava como não condizentes com o modelo de dois estados, e que inicialmente incluímos em nossos cálculos. No entanto, tais sequências aumentavam o χ^2 consistentemente, e portanto também não foram incluídas no conjunto de minimização usado no trabalho final. Todas sequências utilizadas podem ser encontradas no artigo final.

4.4 Procedimento de minimização

O método utilizado para encontrar os parâmetros do modelo PB para a inosina foram os descritos no capítulo 3. Para fazer a minimização nós dividimos as sequências em grupos que contivessem um dado pareamento com a inosina. Por exemplo, um grupo contendo todas sequências que possuíam o par de bases IA (inosina com adenina), outro grupo contendo todas as sequências com IC, e assim por diante. Relembrando, nosso objetivo era determinar os valores ideais para os parâmetros do potencial de Morse (associado às pontes de hidrogênio formadas) e para os parâmetros do potencial harmônico (associado ao empilhamento entre pares de base). O modelo PB em associação com a técnica do índice térmico nos permite fazer predições das temperaturas, e minimizando os quadrados das diferenças entre as temperaturas preditas e as experimentais conseguimos otimizar os parâmetros do modelo (seção 3.4, página 25).

Dividimos o método de minimização em duas etapas, para cada conjunto (IA, IC, etc), descritas a seguir:

Minimização inicial O valor inicial do parâmetro era definido diferentemente para cada rodada de minimização. Para os parâmetros de Morse, o valor inicial era randomicamente escolhido dentro do intervalo 16.5 meV a 49.5 meV (que corresponde a uma variação de $\pm 50\%$ a partir do valor padrão do pareamento AT). Para os parâmetros de empilhamento, de forma análoga, o valor inicial era escolhido no intervalo entre 1.25 meV/nm^2 e 3.75 meV/nm^2. Nesta etapa, após definidos os valores iniciais, todos os parâmetros de empilhamento e parâmetros Morse podiam variar simultaneamente. Durante a minimização os valores dos parâmetros sofriam modificações gradativas, até atingirem valores finais que ofereciam o melhor ajuste aos dados experimentais. O processo foi repetido mil vezes, o que demorou cerca de 600 h de processamento em nossos processadores de 1.6 GHz. O valor resultante de cada parâmetro foi a média

	IA	IC	IG	IT	II
D	31(4)	34(2)	38(2)	29(5)	24(5)
λ	2.9(9)	3.2(2)	3.0(3)	4.9(2)	3.8(1)

Tabela 4.1
Parâmetros do potencial de Morse da inosina. Os potenciais de Morse D são dados em meV e λ é dado em 10^{-2} nm. Em parênteses temos os desvios padrão, obtidos pela última rodada de minimização.

dos 1000 valores obtidos em cada uma destas minimizações.

Minimização final Os valores dos parâmetros iniciais desta etapa eram os valores finais da etapa anterior. Agora, ao invés de variar os parâmetros, iríamos variar o conjunto de dados experimentais. Para um dado conjunto de temperaturas de desnaturação de um tipo de sequência, ou seja IC, variamos as temperaturas de acordo com o erro experimental de 0.3°C reportado por Watkins e SantaLucia [74]. Assim, pudemos estimar quão dependentes são nossos parâmetros do erro experimental envolvido nas medidas de desnaturação. Novamente, todos parâmetros podem variar durante esta etapa, com o objetivo de obter o melhor conjunto de parâmetros possível. Após esta etapa obtivemos o valor ótimo de cada parâmetro, e seu desvio padrão.

Com os resultados na minimização, usamos a média obtida das 1000 rodadas e usamos o desvio padrão como estimativa do efeito da incerteza experimental sobre os valores calculados.

4.5 Resultados para os parâmetros de Morse

Os valores obtidos para os parâmetros do potencial de Morse são mostrados na tabela 4.1. A título de comparação podemos utilizar o valor do potencial de Morse do pareamento AT que, canonicamente, faz duas ligações de hidrogênio com um potencial de Morse de $D_{AT} = 33$ meV [59]. Observando os valores de D na tabela 4.1 percebe-se que todos eles, exceto II, são indicativos de pareamentos com a formação de duas pontes de hidrogênio entre as bases. A capacidade de nosso modelo mesoscópico detectar isso é encorajador, visto que diversos outros estudos foram feitos, para cada pareamento, que corroboram nossos resultados [78, 82–84]. O pareamento II se apresenta mais fraco, o que nos leva a suposição de um encaixe não padrão entre as bases de tal pareamento. Os valores de λ também estão de acordo com o valor aproximado do $\lambda_{AT} = 2.5$ meV/nm^2, com a exceção dos pares IT e II, onde os valores de λ indicam um pequeno alargamento no potencial.

Faixa	IA		IC		I.G		IT		I.I	
	ATpAI	0.0002*			ATpGI	0.013(5)	ATpIT	0.24(8)	IIpAT	0.09(3)
	AIpAT	0.20(6)			IGpAT	0.043(9)	ATpTI	0.6(2)	ATpII	0.8(2)
	IApAT	0.9(2)			GIpAT	0.12(2)	ITpAT	0.9(2)		
< 1					CGpIG	0.69(7)				
≥ 1			CGpCI	1.30(9)			TApIT	1.5(3)		
			CIpAT	1.8(1)						
			CIpCG	1.57(1)						
≥ 2	AIpTA	2.2(3)	ATpCI	2.2(1)	CGpGI	2.0(2)	GCpIT	2.7(4)	CGpII	2.6(4)
	AIpGC	2.2(3)	ATpIC	2.6(2)			CGpTI	2.8(4)		
			ICpAT	2.9(2)						
≥ 3	CGpAI	3.4(3)	GCpCI	3.4(2)						
< 4			CGpIC	3.8(2)						
≥ 4	AIpCG	5.1(4)			GCpIG	4.6(3)	CGpIT	5.9(4)	GCpII	4.6(6)
	GCpAI	7.0(5)			ATpIG	5.8(4)	GCpTI	6.9(4)		
					GCpGI	12.5(4)				

*o desvio padrão para este parâmetro é 0.004 eV/nm^2

Tabela 4.2
Parâmetros de empilhamento k em eV/nm^2. Em parênteses estão os desvio padrão na forma de incerteza de dados compacta. A tabela é dividida em três partes, com a seção central representando os parâmetros de empilhamento que estão na mesma faixa de valores em que se encontram os parâmetros dos pares-de-bases Watson-Crick [59].

4.6 Resultados para os parâmetros de empilhamento

Na tabela 4.2 temos os valores encontrados para os empilhamentos entre cada tipo de vizinhança dos "mismatches" de inosina. As duas linhas horizontais na tabela delimitam a faixa de valores destes parâmetros que é aproximada à dos parâmetros de empilhamento dos pares-de-base exclusivamente canônicos. Abaixo e acima, respectivamente, estão os valores inferiores e superiores à estes limites. Optamos por organizar os resultados assim para melhor visualização de quais parâmetros possuem comportamento usual e quais não possuem. Tal distinção nos permite inferir informações sobre a disposição espacial dos pareamentos da inosina com aquela dada base, e sua interação com seus vizinhos na fita de DNA.

Observando-se os valores da tabela 4.2 podemos notar que o pareamento de inosina que mais se desvia da faixa de comportamento padrão para o DNA é o pareamento IG. Isso nos indica que ele se pareia de modo a distorcer a hélice dupla na região em que ele se encontra como vemos na figura 4.2c. Isto é semelhante ao que foi encontrado por Rutledge et al. [31] que apresenta em seu trabalho, através de cálculos de DFT, em que o ângulo de pareamento ideal para inosina e guanina é tal que o esqueleto de açúcar-fosfato da fita teria de ficar rotacionado para conformar este par.

Por outro lado, ao verificar os valores dos parâmetros de empilhamento do par IC constata-se que eles são os que mais se encaixam aos padrões dos parâmetros de empilhamento canônicos de

Dimer sequence	ΔG°_{37} (kcal/mol)	ΔH° (kcal/mol)	ΔS° (cal/K mol)
I·C pairs			
AI/TC	−0.96 ± 0.06	−8.9 ± 1.4	−25.5 ± 1.4
TI/AC	−0.46 ± 0.06	−5.9 ± 1.5	−17.4 ±1.5
AC/TI	−0.89 ± 0.06	−8.8 ± 1.4	−25.4 ± 1.4
TC/AI	−0.59 ± 0.07	−4.9 ± 1.6	−13.9 ± 1.6
CI/GC	−1.14 ± 0.08	−5.4 ± 1.6	−13.7 ± 1.6
GI/CC	−0.86 ± 0.08	−6.8 ± 1.6	−19.1 ± 1.6
CC/GI	−0.88 ± 0.09	−8.3 ± 1.7	−23.8 ± 1.7
GC/CI	−1.07 ± 0.08	−5.0 ± 1.7	−12.6 ± 1.7
I·A pairs			
AI/TA	−0.51 ± 0.06	−8.3 ± 1.5	−25.0 ± 1.5
TI/AA	0.09 ± 0.06	−3.4 ± 1.4	−11.2 ± 1.4
AA/TI	0.12 ± 0.07	−0.7 ± 1.7	−2.6 ± 1.7
TA/AI	0.12 ± 0.06	−1.3 ± 1.4	−4.6 ± 1.4
CI/GA	−0.18 ± 0.08	2.6 ± 1.6	8.9 ± 1.6
GI/CA	−1.24 ± 0.08	−7.8 ± 1.7	−21.1 ± 1.7
CA/GI	−0.77 ± 0.07	−7.0 ± 1.6	−20.0 ± 1.6
GA/CI	−1.33 ± 0.08	−7.6 ± 1.6	−20.2 ± 1.6
I·T pairs			
AI/TT	0.71 ± 0.06	0.49 ± 1.5	−0.7 ± 1.4
TI/AT	0.36 ± 0.06	−6.5 ± 1.5	−22.0 ± 1.3
AT/TI	0.22 ± 0.07	−5.6 ± 1.7	−18.7 ± 1.6
TT/AI	0.54 ± 0.05	−0.8 ± 1.4	−4.3 ± 1.4
CI/GT	−0.26 ± 0.07	−1.0 ± 1.5	−2.4 ± 1.5
GI/CT	−0.19 ± 0.08	−3.5 ± 1.8	−10.6 ± 1.8
CT/GI	0.41 ± 0.07	0.1 ± 1.5	−1.0 ± 1.5
GT/CI	−0.54 ± 0.08	−4.3 ± 1.8	−12.1 ± 1.7
I·G pairs			
AI/TG	0.02 ± 0.06	−4.9 ± 1.6	−15.8 ± 1.6
TI/AG	0.76 ± 0.06	−1.9 ± 1.5	−8.5 ± 1.5
AG/TI	0.65 ± 0.07	0.1 ± 1.7	−1.8 ± 1.7
TG/AI	0.70 ± 0.06	1.0 ± 1.4	1.0 ± 1.4
CI/GG	0.47 ± 0.10	7.1 ± 2.3	21.3 ± 2.3
GI/CG	−0.10 ± 0.07	−1.1± 1.5	−3.2 ± 1.5
CG/GI	0.54 ± 0.09	5.8 ± 2.0	16.9 ± 2.0
GG/CI	−0.74 ± 0.08	−7.6 ± 1.8	−22.0 ± 1.8
I·I pairs			
AI/TI	0.40 ± 0.05	−3.3 ± 1.2	−11.9 ± 3.9
TI/AI	0.81 ± 0.05	0.1 ± 1.4	−2.3 ± 4.3
CI/GI	0.36 ± 0.06	1.3 ± 1.4	3.0 ± 4.2
GI/CI	−0.09 ± 0.06	−0.5 ± 1.2	−1.3 ± 3.7

Tabela 4.3
Parâmetros de próximos-vizinhos para os diversos pareamentos de inosina obtidos por Watkins e SantaLucia [74]. São mostrados a energia livre de Gibbs em 37°C, a variação de entalpia ΔH e entropia ΔS. A correspondência da notação de próximos vizinhos de adotada por Watkins e SantaLucia com a nossa é a seguinte: AB/CD é um par de base AC seguido de BD, ou seja corresponde ao nosso ACpBD. Por exemplo AI/TC corresponde a ATpIC e AI/TG corresponde a ATpIG. Tabela da referência 74.

DNA. Isto se deve ao fato de que a inosina é estruturalmente mais semelhante à guanina do que às outras bases como se vê na figura 4.1. Desta maneira, este pareamento é o que mais exibirá comportamentos similar ao do pareamentos canônico de CG.

Como vimos, exceto pelo par IC, os demais apresentam variações significativas no empilhamento quando comparados às bases canônicas AT e CG em DNA. Estas constantes de empilhamento são do tipo próximos-vizinhos, como mostramos na seção 3.3, página 21 e na equação (3.34). Então podemos perguntar se os valores de parâmetros de próximos-vizinhos para o pareamento de inosina também diferem tanto assim dos valores de DNA. Na tabela 4.3 mostramos os valores de próximos-vizinhos obtidos por Watkins e SantaLucia [74]. Observe que apenas IC não possui nenhum valor positivo (desestabilizante). Para o DNA, todos os parâmetros são sempre negativos (estabilizante) como vimos na tabela 2.1. Ou seja, observamos sim este tipo de diferença nas caraterísticas de próximos-vizinhos, porém no nosso caso é possível correlacionar isto também com os aspectos estruturais.

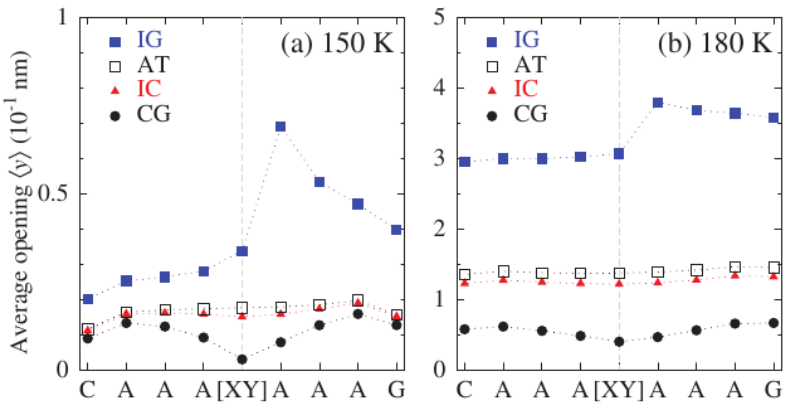


Figura 4.3
Variação do perfil de abertura com a mudança do par de inosina. Os cálculos para obtenção dos painéis (a) e (b) diferem apenas na temperatura em que foram feitos, indicadas no canto superior direito de cada painel. Para comparação, também usamos os pares de base canônicos AT e CG na posição central.

4.7 Análise dos perfis de abertura de sequências contendo inosina

Uma vez estabelecidos os parâmetros, nós podemos calcular os perfís de abertura média $\langle y_i \rangle$ usando a equação (3.32). Note que o cáculo de $\langle y_i \rangle$ é independente da regressão das temperaturas dada pela equação (3.39), ou seja a temperatura T usada na equação (3.32) não tem nenhuma relação com a temperatura de desnaturação (veja também a equação (3.24), . Um valor baixo de $\langle y_i \rangle$, na ordem de 1 nm, indica um DNA hibridizado, valores acima deste indicam DNA desnaturado. Um problema que ocorre é que para sequências muito curtas, em temperaturas na mesma faixa das temperaturas de desnaturação o valor de $\langle y_i \rangle$ é tipicamente maior que 10 nm, ou seja, o DNA já estaria desnaturado. Para obter perfis de abertura de uma fita hibridizada precisamos reduzir substancialmente a temperatura do cálculo, muitas vezes na faixa de 100 K quando a sequência é menor que 20 bp. Este é uma limitação bem conhecida do modelo [52, 85]. Apesar desta limitação, os perfis de abertura permitem obter uma visualização das influências dos parâmetros na molécula de DNA. Na figura 4.3 temos um exemplo de perfil de abertura com uma sequência formada por pares canônicos (onde omite-se o par correspondente) a não ser na posição central, indicada por XY. Nesta posição temos diversos pares de base diferentes para comparação do efeito que eles causam na sequência. Nota-se que o par que mais estabiliza a sequência é o par CG, o que era de se esperar já que possui o maior valor de D no potencial de Morse, correspondente a três ligações de hidrogênio. Os pares AT e IC estabilizam de forma praticamente idêntica a sequência, como já havia sido observado na seção 4.6. O pareamento IG, por ter parâmetros de empilhamento mais fracos no contexto apresentado, é o que mais desestabiliza a sequência. Na figura 4.4 mostramos o efeito do par IA onde se observa uma descontinuidade do perfil de abertura a partir da posição em que inosina está presente, para o pareamento em que o empilhamento é mais fraco de acordo com a tabela 4.2. Note que ao aumentar a temperatura, mostrado na figura 4.4b, o lado esquerdo tem

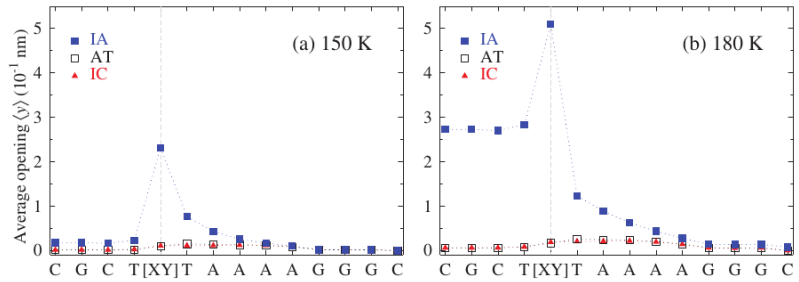


Figura 4.4
Demonstração dos efeitos causados pelo empilhamento fraco entre o pareamento da inosina e o par de base vizinho em duas temperaturas diferentes. O pareamento AT é colocado como referência.

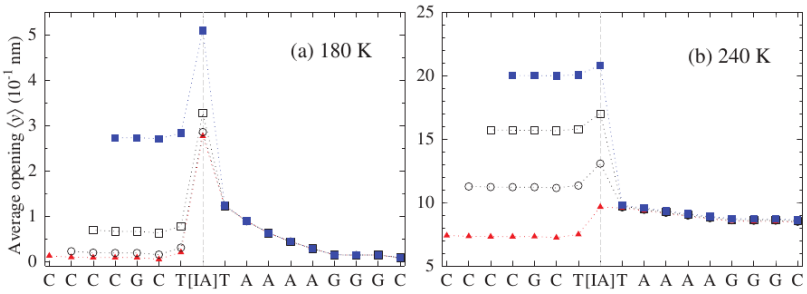


Figura 4.5
Distorção causada pela inosina em relação ao tamanho da fita. Cada símbolo indica uma sequência diferente, todas formadas por um trecho em comum ao qual foi adicionado, consecutivamente, um par CG adicional após a posição de pareamento da inosina.

um aumento significativo do $\langle y \rangle$ que não é acompanhado pelo lado direito. Assim, efetivamente o valor extremamente baixo de ATpAI, tabela 4.2, causa uma perda de cooperatividade entre os dois lados da sequência de DNA. A figura 4.5 mostra este efeito variando o tamanho do lado esquerdo da sequência, na medida em que aumentamos a quantidade de CGs no lado esquerdo diminuímos o efeito da perda de cooperatividade.

4.8 Conclusões

Com a parametrização da inosina em DNA conseguimos reproduzir o fato bem conhecido que esta base se liga com qualquer outra por duas ligações de hidrogênio. De fato, os potenciais de Morse para os quatro pareamentos são muito próximos ao de AT. Outro resultado interessante é que mostramos que IC, que tem uma estrutura muito parecida com GC, tem uma interação de empilhamento na mesma faixa de DNA normal. Já para IA, a única com configuração syn-anti, os valores de empilhamento assumem valores extremos chegando a causar uma ruptura na cooperatividade como vista nos perfis de abertura média. Assim, confirmamos com este trabalho a capacidade do modelo PB, e da técnica de parametrização por equivalência térmica, de não apenas obter o número de ligações de hidrogênio consistentes com a literatura, mas também de nos fornecer informações estruturais sobre a conformação dos pares de bases na fita.

Capítulo 5

Estrutra e funcionalidades do RNA

Ao longo desta tese fazemos continuamente referências à diversos tipos de ácidos nucleicos, citando sua importância biológica e tecnológica. Neste capítulo tentaremos expor de forma breve cada um dos tipos de RNA de funcionalidades mais estudadas. Sobre RNA iremos tratar especificamente dos rRNA, tRNA, mRNA, siRNA, miRNA.

5.1 rRNA — RNA ribossomal

O RNA é o material predominante na estrutura dos ribossomos, organelas presentes no corpo humano e ilustradas na figura 5.1. O ribossomo é composto por aproximadamente 60% de rRNA e 40% de proteína (em termos de massa). As duas maiores sequências de rRNA que formam o ribossomo são as sub-unidade maior (*large sub-unit*, LSU), e a sub-unidade menor, sub-unidade menor (*small sub-unit*, SSU) (*small sub-unit*). Na figura 5.1 também estão evidentes as duas sub-unidades.

Detalhes à respeito da estrutura molecular de cada tipo de rRNA serão dados à seguir. A SSU

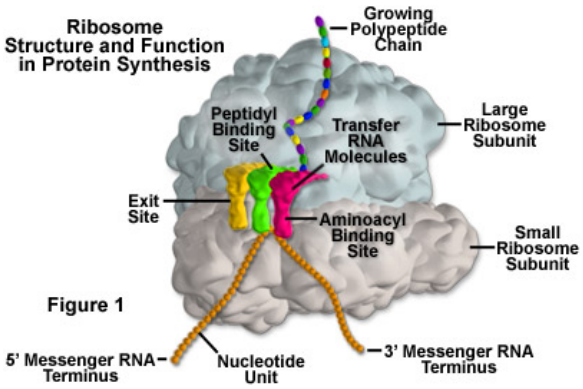


Figura 5.1
O RNA ribossomal, sub-unidades maior e menor. Figura da referência 86.

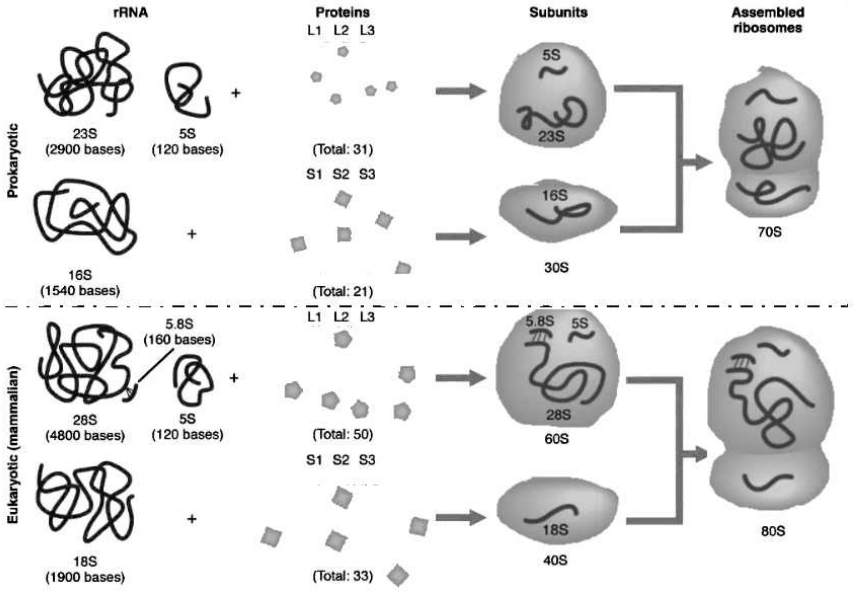


Figura 5.2
O rRNA de procariotos (metade superior) e eucariotos (metade inferior) no processo de formação de um ribossomo. Figura da referência 90.

possui de 1000 a 2000 nucleotídeos de tamanho [87, 88]. As sequências deste subtipo de rRNA são usadas para estudos de relacões evolutivas entre as espécies [89], já que elas são encontradas em uma quantidade enorme de formas de vida. Em procariotos, por exemplo, temos o ribossomo 70S, formado pelo complexo 30S que consiste no rRNA 16S acompanhdo de 19 proteínas. É justamente o 16S rRNA que é usado para reconstruir as filogenias e árvores genéticas, já que essa região do gene apresenta uma lenta taxa evolutiva. A figura 5.2 ilustra a montagem dos ribossomos mais comuns em procariotos e eucariotos.

O LSU, por sua vez, possui na faixa de 1600 a 5000 nucleotídeos em sua sequência [87, 91]. Ele é responsável por agir como um catalisador da formação de ligações peptídicas (síntese por desidratação que ocorre entre moléculas de aminoácidos). Assim como a SSU, a LSU é usada para mapeamento filogenético[92], por estar presente em praticamente todo ser vivo, e também apresentar uma lenta taxa evolutiva de transformação. Nos eucariotos o ribossomo 80S é composto pelo LSU[1] 60S e mais 47 proteínas [93]. A sub-unidade maior pode ser vista, também, na figura 5.2.

5.2 tRNA — RNA de transferência

Uma das classes de macromoléculas mais estudadas é a dos RNAs de transferência. Elas possuem de 76 a 90 nucleotídeos [94]. Tais moléculas são essenciais no processo de tradução do código genético para obtenção de proteínas, pois elas carregam os aminoácidos até os códons correspon-

[1] Este por sua vez é composto pelas sequências rRNA 5S, 28S e 5.8S [93].

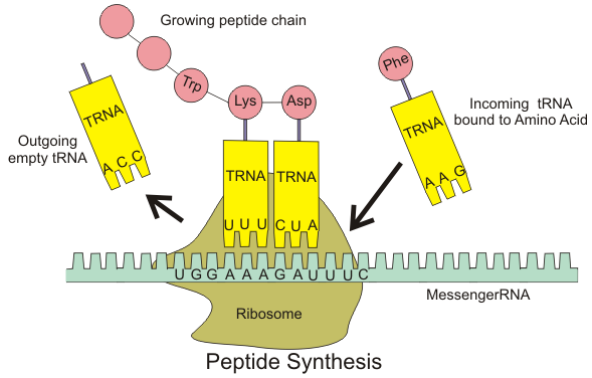


Figura 5.3
O papel do RNA de transferência é levar o aminoácido até o local apropriado para a tradução ocorrer [94]. Figura da referência 86.

	U	C	A	G	
U	UUU Fenilalanina	UCU Serina	UAU Tirosina	UGU Cisteína	U
	UUC Fenilalanina	UCC Serina	UAC Tirosina	UGC Cisteína	C
	UUA Leucina	UCA Serina	UAA Parada	UGA Parada	A
	UUG Leucina	UCG Serina	UAG Parada	UGG Triptofano	G
C	CUU Leucina	CCU Prolina	CAU Histidina	CGU Arginina	U
	CUC Leucina	CCC Prolina	CAC Histidina	CGC Arginina	C
	CUA Leucina	CCA Prolina	CAA Glutamina	CGA Arginina	A
	CUG Leucina	CCG Prolina	CAG Glutamina	CGG Arginina	G
A	AUU Isoleucina	ACU Treonina	AAU Asparagina	AGU Serina	U
	AUC Isoleucina	ACC Treonina	AAC Asparagina	AGC Serina	C
	AUA Isoleucina	ACA Treonina	AAA Lisina	AGA Arginina	A
	AUG Metionina	ACG Treonina	AAG Lisina	AGG Arginina	G
G	GUU Valina	GCU Alanina	GAU Aspartato	GGU Glicina	U
	GUC Valina	GCC Alanina	GAC Aspartato	GGC Glicina	C
	GUA Valina	GCA Alanina	GAA Glutamato	GGA Glicina	A
	GUG Valina	GCG Alanina	GAG Glutamato	GGG Glicina	G

Tabela 5.1
Associação entre códons e aminoácidos. Tabela da referência 95.

dentes que estão dispostos na fita de mRNA, discutido mais adiante na seção 5.3.

O tRNA possui uma estrutura molecular que o permite conectar-se com um aminoácido em uma de suas extremidades, e fazer um pareamento de três pares de base com um outro trecho de RNA (especificamente, com um códon do mRNA, discutido na seção 5.3). A estrutura dele é parecida com o que chamamos de folha de trevo, e pode ser visto na figura 5.4. Repare no loop (a estrutura em forma de laço) na parte inferior da figura. Falaremos mais de loops e outras estruturas morfológicas no capítulo 6. Neste loop de tRNA, especificamente, temos o anticodon GAA — a parte que irá se conectar com o codon CUU presente no mRNA. O aminoácido correspondente ao CUU está na tabela 5.1.

Sobre o tRNA e sua estrutura molecular, o capítulo 6 irá discutir as três classificações estruturais que descrevem o RNA: estruturas primárias, secundárias e terciárias (de forma grosseira, respectivamente: informação da sequência, informação dos pareamentos, ângulos, rigidez e torções tridimensionais).

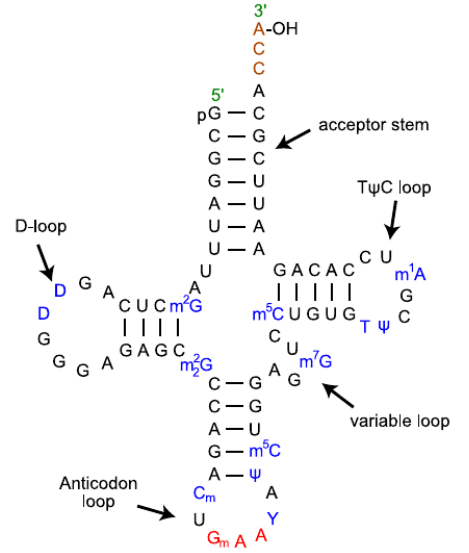


Figura 5.4
O RNA de transferência possui uma estrutura em formato de trevo para cumprir seu papel no dogma central da biologia molecular. Figura da referência 96.

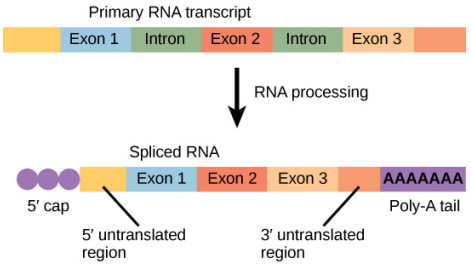


Figura 5.5
O RNA mensageiro passa por um tratamento em que algumas das regiões são removidas (introns) e outras permanecem na sequência (exons). Figura da referência 98..

5.3 mRNA — RNA mensageiro

O RNA mensageiro (*messenger RNA*, mRNA), é uma sequência de RNA que carrega a informação contida no trecho de DNA do qual ele é transcrito. É o mRNA que será lido para que ocorra a síntese proteica correspondente aos genes originais (lidos a partir do trecho de DNA). Após a transcrição (de DNA para mRNA a sequência de nucleotídeos recebe um tratamento que separa a sequência em diferentes partes: introns e exons. Introns são partes que são removidas da sequência de RNA, exons são as que permanecem e são reorganizadas. Os introns não são traduzidos em proteínas, mas possuem diversas funções relevantes na síntese proteica [94, 97]. Os exons são a parte do mRNA que será traduzido para formação das proteínas daquele ser vivo. A figura 5.5 ilustra o processo de transcrição a partir do DNA, remoção dos introns, e reajuste da posição dos exons.

Os exons são as regiões onde teremos conjuntos de códons consecutivos. Um códon é uma sequência de 3 nucleotídeos de mRNA. Cada códon irá se parear com um anticodon (veja também a seção 5.2) específico do tRNA que indicará qual o próximo aminoácido que deve ser considerado na síntese proteica. A relação códon-aminoácido é apresentada tabela 5.1. Um único mRNA, referente a um dado trecho de DNA transcrito, pode ser preparado de mais de uma maneira pelas enzimas

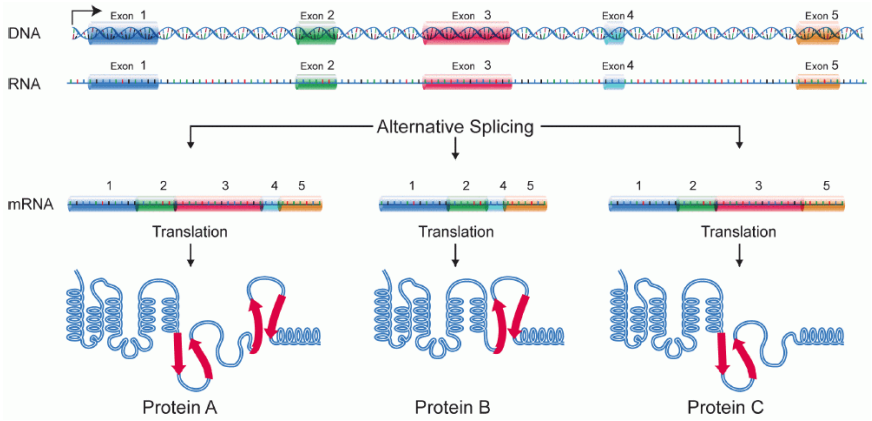


Figura 5.6
Uma mesma sequência de RNA mensageiro pode sofrer diferentes processos de maturação, vindo a ser traduzida em mais de um tipo de proteína (A,B e C). Figura extraída de [98].

que regulam o processo de maturação (remoção dos introns e preparação para síntese proteica). A figura 5.6 mostra três possíveis proteínas (A,B e C) que poderiam ter sido obtidas a partir da mesma sequência original de mRNA não tratado. Como exemplo, temos uma sequência fictícia AGUC. Removendo um único nucleotídeo (i.e. "Intron" do nosso exemplo), teríamos os seguintes resultados (i.e. "Exons"): AGU, GUC (sem reorganização, apenas removendo um nucleotídeo) e AGC (com remoção de um nucleotídeo, seguido de junção das partes restantes).

É importante notar que na preparação do mRNA para ser traduzido ele irá receber sequências de início e de terminação: a Cap $5'$ e a cauda poli-A. Elas protegem, ao menos por um tempo, o mRNA da degradação inevitável que ele irá sofrer no interior celular, onde seu tempo de meia-vida é cerca de 7 a 10 horas [99][2]. Ambas estruturas também possuem papel fundamental na interação do mRNA com o ribossomo: elas ajudam na identificação da sequência a ser lida, como posições de início e de término da leitura [94].

5.4 miRNA, siRNA — RNAs que regulam outros RNAs

O RNA curto de interferência (*small interfering RNA*, siRNA) é uma classe de moléculas de RNA que se organizam em forma de hélice dupla, possuem tamanho de 20 a 25 pares de base (similarmente ao miRNA). Eles degradam o mRNA após a transcrição, o que evita que ocorra a tradução. Discutiremos em breve como funciona o processo de interferência.

O microRNA (*microRNA*, miRNA), é uma molécula de RNA de aproximadamente 22 nucleotídeos que possui um papel na regulação de expressão gênica. Em outras palavras, ele regula quanta proteína será produzida de forma direta ou indireta [100, 101]. Ele assume a forma de um *hair-*

[2]Medidas de mRNA de mamíferos.

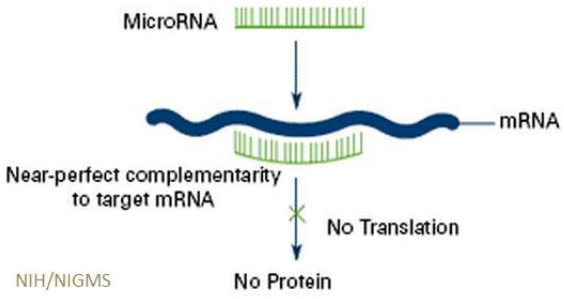


Figura 5.7
O processo de regulação gênica realizado pelo miRNA. Figura da referência 102.

pin (veja também o capítulo 6). Um pareamento aproximadamente perfeito entre o miRNA e um trecho do mRNA que seria expressado irá provocar a degradação daquele mRNA. Este é um dos mecanismos de regulação gênica. O outro é quando ocorre um pareamento imperfeito: a tradução continuará ocorrendo, porém em uma taxa reduzida devido à presença do miRNA, e a degradação será acelerada (pela quebra da cadeia poli-A mencionada na seção 5.3). A figura 5.7 representa a regulação gênica provocada pelo miRNA.

5.5 Conclusão

Neste capítulo revimos as principais estruturas do RNA com suas funções biológicas. O que todas tem em comum é a formação de estruturas complexas em forma de laços. No próximo capítulo veremos em mais detalhe como estas estruturas podem ser calculadas usando conceitos de programação dinâmica e o modelo de próximos vizinhos (*nearest-neighbour model*, NN) que vimos anteriormente no capítulo 2.

Capítulo 6

Estrutura Secundária de RNA

Neste capítulo iniciamos com uma motivação, na seção 6.1 que demonstre ao leitor o porque de termos iniciado nossas investigações no campo de estruturas secundárias de RNA. Em seguida, apresentamos algumas definições na seção 6.2 para familiarizar o leitor com os termos técnicos utilizados na área de análise de estrutura secundária. Na seção 6.3 é apresentado o algoritmo de previsão de estruturas secundárias das sequências de RNA e discutiremos o software que utilizamos para cálculo destas previsões. As estruturas subóptimas são abordadas na seção 6.4. Na seção 6.5.3 apresentamos as métricas que são usadas para discutir a qualidade das predições. Temos as análises comparativas entre subóptimos e a randomização de dados experimentais em 6.5, e, por fim, as seções 6.5.2 e 6.5.5 apresentam uma análise sobre como melhorar os parâmetros termodinâmicos utilizados na predição da estrutura secundária.

6.1 Motivação

Um grupo de pesquisadores da UFRGS estava investigando um grupo de pacientes com câncer hereditário na região da tireoide como parte da tese de doutoramento de Lucieli Ceolin, na época aluna sob supervisão de Ana Luiza Maia. Em suas análises bioquímicas das sequências de RNA dos pacientes com carcinoma medular da tireoide (*medullary thyroid carcinoma*, MTC) encontraram um problema na determinação da estrutura secundária relacionado ao mRNA. No caso, a sequência de estudo era pertencente ao cromossomo humano 10, e codifica uma proteína de membrana de alguns tipos celulares presentes na tireoide e nas glândulas supra renais.

A análise das sequências extraídas dos pacientes apontava que haviam algumas substituições de bases pontuais na sequências dos mesmos, porém não havia diferenças nos aminoácidos utilizados

na confecção das proteínas deles em relação à pessoas sem a doença. Quando há alteração de um único nucleotído na sequência sem haver alteração da proteína dizemos que temos um caso de polimorfismo de nucleotídeo único (*single nucleotide polimorfism*, SNP). Isto é possível graças ao fato de possuirmos uma relação degenerada entre códons e aminoácidos.

Como tentativa de explicar o fenômeno, nossos esforços conjuntos se voltaram à análise da estrutura secundária destas sequências de RNA. Vimos que não havia diferenças significativas entre as estruturas, o que era parcialmente esperado, dado que cada sequência composta de $\approx$5000 bp diferia da sequência selvagem por menos de 5 bp. Na tentativa de descobrir mais sobre a estrutura das sequências começamos a investigar o significado das estruturas suboptimas, discutidas em 6.4. Supostamente, no conjunto de tais estruturas diferentes que um ensemble subóptimo pode nos fornecer, teríamos a informação estrutural da sequência de RNA no contexto de flutuações ao redor do ponto de equilíbrio termodinâmico, e de pequenas variações ou imprecisões inerentes ao ato de medida. A íntegra do artigo está no apêndice E.

Porém, após a publicação do artigo algumas perguntas surgiram:

1. Quão confiável é uma predição de estrutura secundária de uma sequência de 5000 bp?

2. Os subóptimos de fato representam conformações possíveis que poderíamos obter a partir da incerteza dos dados experimentais utilizados na predição?

Os trabalhos que seguiram, e cujos resultados serão expostos neste capítulo, tem o intuito de responder a estas perguntas.

6.2 Elementos das estruturas secundárias

No estudo da estrutura dos ácidos nucleicos nós utilizamos uma classificação hierárquica das informações necessárias para a descrição da molécula em questão. A figura 6.1 ilustra o que chamamos de estrutura primária, secundária e terciária. A estrutura primária corresponde à informação mais básica da sequência: quais são os nucleotídeos que a compõe e em qual ordem eles se encontram, podemos pensar nela como uma informação unidimensional (1D). Por exemplo, AAACCCGGG e ACGACGACG são exemplos de estruturas primárias usadas para descrever ácidos nucleicos. Em ambos os casos temos três guaninas, três citosinas e três adeninas. A estrutura primária, no entanto, já evidencia a diferença destas duas sequências no que diz respeito à como tais nucleotídeos estão

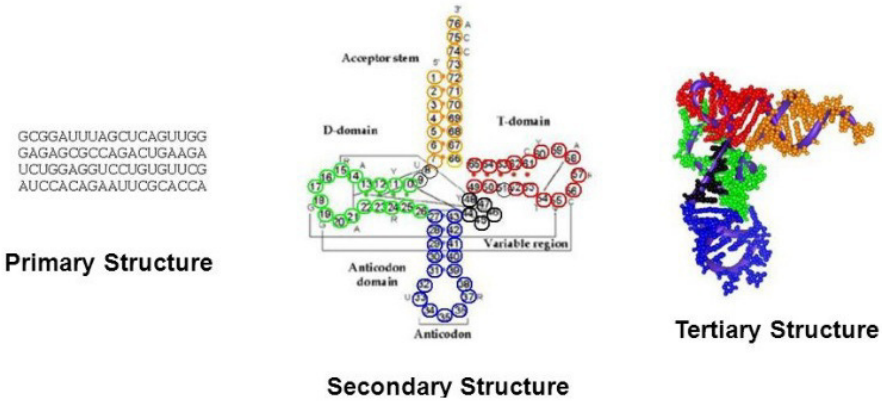


Figura 6.1
A hierarquia da descrição estrutural do RNA, apresentando a estrutura primária, secundária e terciária.

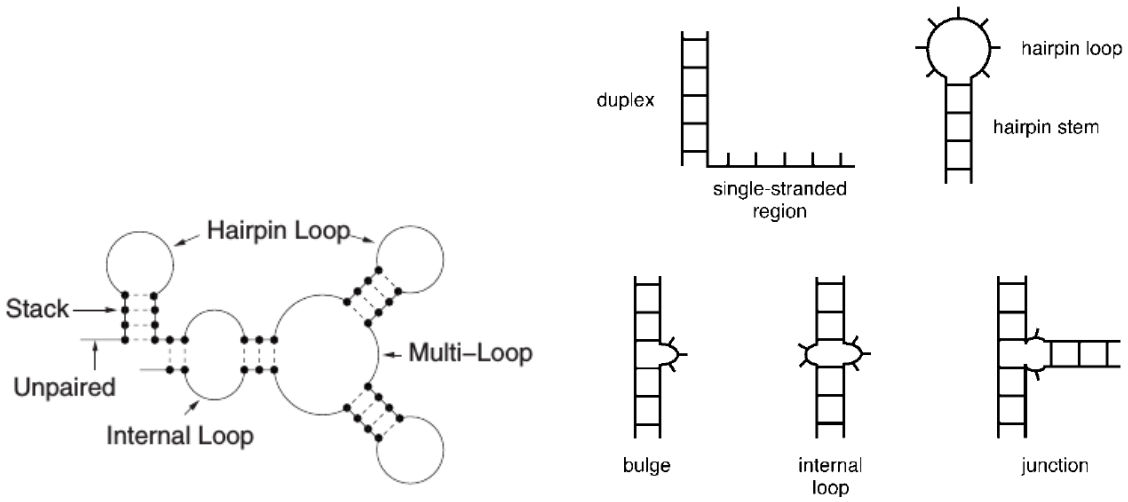


Figura 6.2
Elementos que podem ocorrer na estrutura secundária de RNA. Figuras das referências 18, 103.

organizados ao longo da molécula. Em seguida temos a chamada estrutura secundária: ela nos informa quais bases estão ligadas a quais outras, veja a parte central da figura 6.1, podemos considerá-la como uma representação bidimensional (2D) focando somente nas ligações entre bases. Por fim, temos a estrutura terciária que é a representação tri-dimensional da molécula. Conhecer a estrutura secundária do RNA é de extrema importância para diversas aplicações em biologica molecular.

Neste trabalho vamos nos concentrar nas estruturas secundárias que a molécula de RNA pode formar, e veremos que determinar esta representação simplificada já nos traz informações relevantes sobre a molécula. Há diversos elementos nas estruturas secundárias que são fácilmente identificáveis como podemos ver na figura 6.2. Em particular, o que se convencionou chamar de *stack* é um trecho de fita dupla em geral do tipo A como vimos na introdução, figura 1.4, página 5. O termo *stem* também pode ser usado. A fita simples, com todas bases daquela região sem pareamento com

outras bases recebe o nome de fita simples, ou *single stranded region*, como visto na mesma região da figura. Estas são regiões que ocorrem em meio à duplas hélices, mas que apresentam secções de pares de bases sem pareamento de um lado da fita, ou dos dois lados. Loops são estruturas que podem ter uma quantidade de pares de base não pareados indefinidamente longa, desde que a extremidade deste trecho da fita não pareado volte a se conectar com alguma base próxima à região de onde o segmento teve início. Um *bulge*, ou calombo, é um *loop* de poucas bases (uma, duas, ou no máximo três), que aparece de um único lado, e que retorna à fita principal de forma que a fita oposta da dupla hélice se apresente totalmente pareada. Por fim, temos a estrutura de *hairpin*, ou grampo-de-cabelo, que é quando um único trecho da fita faz um *loop* e retorna dobrando-se sobre si mesmo. A parte que forma o *loop* é denonimada *hairpin loop* e a parte de hélice dupla é chamada de *hairpin stem*.

6.2.1 Representações da estrutura secundária

Uma estrutura secundária S de uma sequência de RNA de comprimento n, é uma rede (ou grafo) com conexões entre os nós (ou vértices). Cada base da sequência de RNA representa um nó e será numerado sequencialmente, iniciando pelo terminal $5'$, formando um conjunto $\{1, \ldots, n\}$ como mostrado no exemplo da figura 6.3a. Cada pareamento da base localizada em i com outra base em j será representado por $(i|j)$, em particular a seguinte condição precisa ser satisfeita

Condição 1. *Cada nó (base) recebe no máximo uma ligação.*

Ou seja, só formamos pares de base ou então a base está despareada. Isto em particular acaba por excluir formações triplexas [104] ou quadruplexas [105] em RNA. A representação dos pareamentos no exemplo da figura 6.3a são os os pares de base $(2|6)$, $(3|5)$ e $(7|10)$. Deste exemplo podemos tirar a primeira propriedade matemática, vamos considerar dois pares de base na estrutura, sejam $(i|j)$ e $(k|l)$, então

Condição 2. *Se $k < j$ então necessariamente $i < k < l < j$.*

No nosso exemplo, $(i = 2|j = 6)$ $(k = 3|l = 5)$ satisfazem esta condição. Podemos entender melhor esta condição olhando para a estrutura linearizada da figura 6.3b onde os pares de base são representados por arcos conectando os nós. Nota-se que nos pares que satisfazem a condição (2) os arcos contornam um ao outro, ou seja o par $(k = 3|l = 5)$ forma um arco interno ao $(i = 2|j = 6)$. A violação da condição (2) leva a formação de pseudo-nós (*pseudo-knots*), como mostrados na

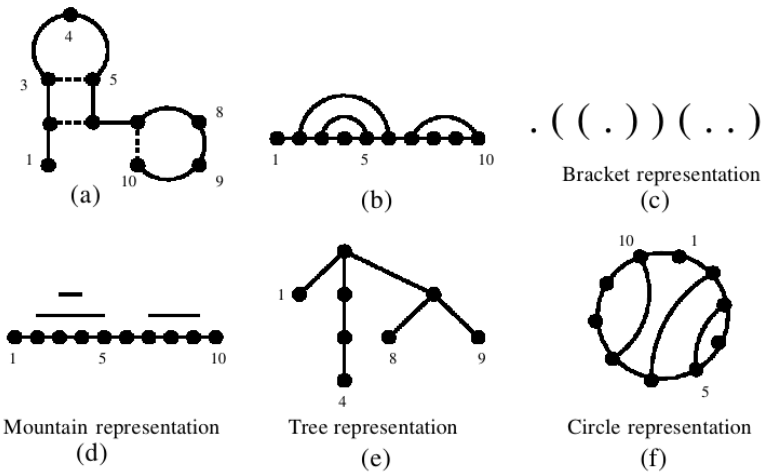


Figura 6.3
Representações da estrutura secundária de RNA. Figura da referência 107.

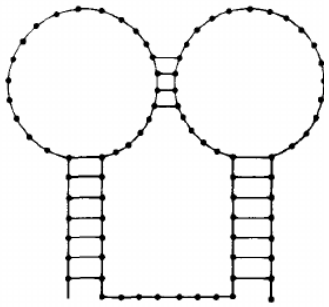


Figura 6.4
Estrutura secundária de RNA com pseudo-nós, neste caso ligações entre os *loops*, que violam a condição (2). Figura da referência 108.

figura 6.4, e faria com que estes arcos se cruzassem. O terceiro par, $(7|10)$, não satisfaz $k < j$ com nenhum outro par, então a condição (2) não se aplica. Uma outra maneira de visualizar a condição (2) é a representação circular introduzida por Nussinov et al. [106] e exemplificado na figura 6.3f. Aqui se alinha a sequência de RNA num círculo numerando as bases em ordem horária, os pares são ligados por arcos e, se a condição (2) é satisfeita, não pode haver cruzamentos. Além das condições 1 e 2 há ainda a condição:

Condição 3. *Não pode have ligações consecutivas do tipo* $(i|i+1)$.

Note que isto implicaria em uma ligação de hidrogênio entre bases empilhadas, o que não seria possível como se pode ver pela figura 1.5, página 5.

Como vimos a estrutura secundária representa uma espécie de visão bidimensional da molécula, no entanto para os algoritmos que determinam tais estruturas ela também pode ser expressa em forma unidimensional. A vantagem da forma unidimensional, uma espécie de informação primária da estrutura secundária é que isto nos permite a aplicação de técnicas de programação dinâmica que são usados no alinhamento de estruturas primárias de DNA [109].

A notação mais usual é a *bracket representation* ou notação *dot-brackets* [107]. Do exemplo anterior a notação *dot-brackets* pode ser vista na figura 6.3c onde cada pareamento $(i|j)$ é substituído

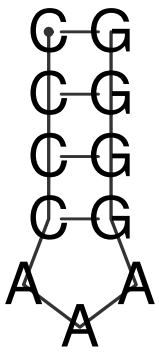


Figura 6.5
Representação da estrutura secundária de menor energia da sequência CCCCAAAGGGG. Figura feita com o software [110].

por um par de colchetes e quando uma base não forma nenhum par usa-se um ponto. Para exemplificar esta notações iremos usar uma única fita de RNA, que irá dobrar sobre si mesma com as bases de uma parte da fita interagindo com as bases de outra parte. Seja esta sequência CCCCAAAGGGG. Suponha que a parte inicial, composta de citosinas, faça pareamento com a parte final, composta de guaninas. A estrutura secundária, na sua representação bidimensional, seria algo semelhante ao exposto na figura 6.5.

Cada uma das notações, naturalmente, deve representar a mesma estrutura que vemos na figura 6.5. Agora veremos como funciona cada uma. Primeiramente a notação *dot-brackets*. Uma base que não esteja pareada com ninguém é representada por um ponto ".", uma base pareada é representada com um conjunto de parênteses que se abrem e fecham "()". O último parênteses aberto faz um par com o primeiro parênteses fechado, e em seguida busca-se o próximo conjunto semelhante. Obtemos assim o seguinte resultado

Estrutura primária:	C	C	C	C	A	A	A	G	G	G	G
dot-brackets:	(	(	(	(	.	.	.	)	)	)	)

onde o código de cor mostra os parênteses correspondentes.

A tabela de conectividade (*connectivity table*, CT) não é propriamente uma representação mas um formato de tabela que é usado em diversos *softwares*. Na tabela 6.1, a primeira linha se inicia com a informação de quantas bases existem na sequência a ser descrita. Após este valor temos a informação do título da estrutura. Em cada linha adicional teremos informações a respeito do pareamento da sequência, organizadas em colunas. A primeira coluna representa o índice n que identifica a posição da base (primeira base, segunda base, etc.). A segunda coluna representa qual base nitrogenada ocupa aquela posição (A, C, G ou U). As colunas seguintes são indexações, terceira coluna é $n-1$ e a quarta coluna é $n+1$ representando os índices dos próximos vizinhos. A quinta coluna mostra o índice n_2 da base que está fazendo pareamento com a base n, quando não há

11		Exemplo-RNA			
n	base	$n-1$	$n+1$	n_2	n
1	C	0	2	11	1
2	C	1	3	10	2
3	C	2	4	9	3
4	C	3	5	8	4
5	A	4	6	0	5
6	A	5	7	0	6
7	A	6	8	0	7
8	G	7	9	4	8
9	G	8	10	3	9
10	G	9	11	2	10
11	G	10	0	1	11

Tabela 6.1
Representação da estrutura secundária da sequência CCCCAAAGGGG na notação tipo CT comumente empregada em *softwares* de predição de estrutura secundária.

pareamento é usado zero. A última coluna é simplesmente o índice n repetido.

6.3 Algoritmos para predição da estrutura secundária

Sendo a forma das estruturas RNA tão intimamente relacionadas à sua função, a predição de tais estruturas se tornou de grande interesse. Prevendo a estrutura proveniente de uma dada sequência, podemos, por exemplo, por análise comparativa, associar a qual grupo funcional ela pertence.

Historicamente, as primeiras tentativas em prever a estrutura secundária usando algoritmos computacionais são de Fresco et al. [112] e Delisi e Crothers [113], fazendo uso da ideia de energia livre de Gibbs, equação (2.13), página 12. Ou seja, determinando a estrutura secundária que assuma a *menor* energia livre e que, portanto, seria a estrutura mais estável do ponto de vista termodinâmico. Tinoco et al. [111] empregam esta mesma ideia, mas também introduzem a utilização de uma matriz de adjacência como mostrado na figura 6.6. A matriz de adjacência é uma ideia oriunda de teoria de redes, ou grafos, que representa as ligações entre os nós em forma matricial. O conceito é empregado em alinhamento de sequências, por exemplo no algoritmo de Needleman e Wunsch [109], onde uma sequência de DNA é comparada com outra, colocando uma nas colunas e outra nas linhas. No caso de estrutura secundária de RNA, a sequência é comparada com ela mesma, por isto ela é colocada tanto nas colunas como nas linhas como mostrado na figura 6.6. O cálculo da matriz se dá por regras de preenchimento, onde um elemento da matriz é calculado baseado nos elementos vizinhos. Hoje em dia, a maioria dos *softwares* se baseia no algoritmo de Nussinov [106, 114, 115] para avaliar as matrizes de adjacência. Este algoritmo implementa as condições (1–3) e a regra de

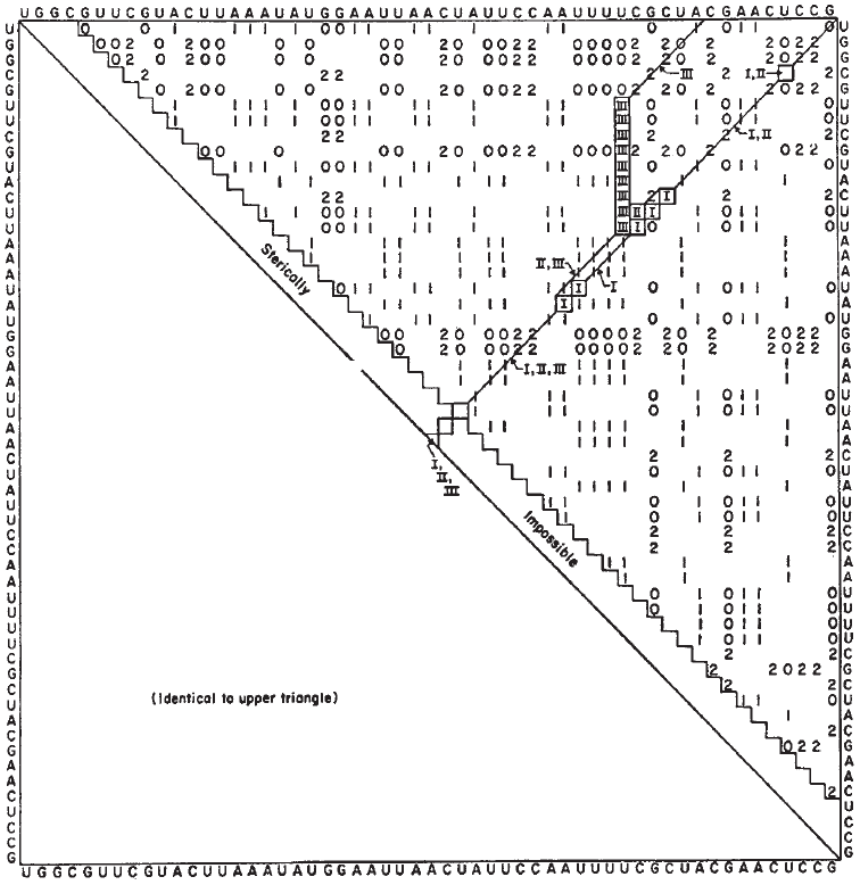


Figura 6.6
Exemplo de matriz de adjacência para determinar a estrutura secundária de RNA. Figura da referência 111.

preenchimento da matriz de adjacência

$$S(i|j) = \max \begin{cases} S(i+1|j-1) + E(i,j), & \text{se forma um par} \quad (i|j) \\ S(i+1|j) \\ S(i|j-1) \\ \max_{i<k<j} S(i|k) + S(k+1|j) \end{cases} \tag{6.1}$$

Onde $E(i,j)$ é a pontuação que se dá quando (i|j) forma uma base. Na sua formulação mais básica $E(i,j) = 1$ [106, 116]. A aplicação sequencial, elemento por elemento, da matriz é conhecido por programação dinâmica. Inicia-se pelo preenchimento da diagonal $S(i|i) = 0$ já que a base não pode pode parear com ela mesma. A diagonal seguinte, $S(i|i+1)$ também tem de ser zero por causa da condição (3). Esta parte, chamada de *initialization*, pode ser vista na figura 6.7. O restante da matriz é preenchida aplicando a equação (6.1), processo chamado de *recursive fill*, sempre na diagonal seguinte $S(i|i+n)$. O último elemento preenchido será sempre $S(1|N)$.

Uma vez preenchida a matriz, a estrutura é determinada pelo processo de *traceback* ou *backtrack*. Enquanto o processo de preenchimento é muito simples, o *traceback* é um pouco mais sutil. Inicia-se o *traceback* pelo último elemento preenchido $S(1|N)$ e procura-se o elemento vizinho de

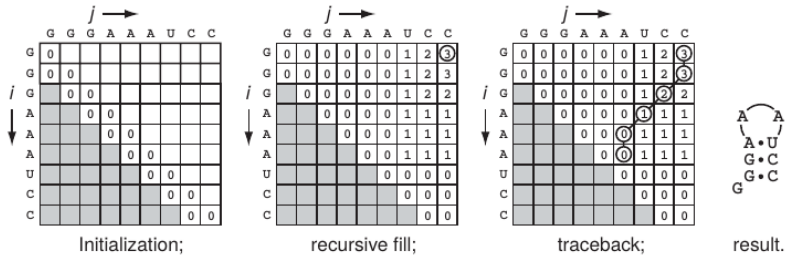


Figura 6.7
Preenchimento da matriz de adjacência. Figura da referência 116.

maior valor e faz-se uma ligação entre esses dois elementos. Se esta ligação se der ao longo diagonal inversa considera-se que há uma estrutura formada. O algoritmo em pseudocódigo é o seguinte [117]

```
−push(1,N);
−pop(i,j).
−if i < j continue;
  else if S(i+1,j)=S(i,j) push(i+1,j);
  else if S(i,j−1)=S(i,j) push(i,j−1);
  else if S(i+1,j−1)+E(i,j)=S(i,j):
     −record i,j base pair.
     −push(i+1,j−1).
  else for k=i+1 to j−1: if S(i,k)+S(k+1,j)=S(i,j):
     −push(k+1,j).
     −push(i,k).
     −break.
```

A regra importante deste pseudocódigo é quando S(i+1,j−1)+E(i,j) for igual a S(i,j), note que $i+1, i-1$ é precisamente na diagonal inversa da posição i, j que mencionamos. A figura 6.7 mostra um exemplo de *traceback* quando se aplica este pseudocódigo a uma sequência GGGAAAUCC, e usando $E(i,j) = 1$ sempre que i e j formarem um par, teremos uma estrutura do tipo

```
.((.(.)))
```

mas note que um pequeno rearranjo nos testes do pseudocódigo do tipo

```
  else if S(i+1,j−1)+E(i,j)=S(i,j):
  else if S(i+1,j)=S(i,j) push(i+1,j);
  else if S(i,j−1)=S(i,j) push(i,j−1);
```

resulta em

```
((.(..)))
```

que é uma estrutura também com três pares de base que é tão válida quanto a anterior. Isto ilustra que o *traceback* tem mais de um resultado possível, mesmo para a regra mais elementar usando $E(i,j) = 1$.

Claramente, a pontuação $E(i,j) = 1$ (sempre que i e j formarem um par) não representa a termodinâmica da molécula como vimos anteriormente no capítulo 2. Waterman e Smith [118] usam

as energias livre de Gibbs, equação (2.13), página 12, para a função $E(i,j)$ e provam que estruturas secundárias calculadas desta maneira são estruturas de energia livre mais baixa, ou seja energia livre mínima (*minimum free energy*, MFE)[1]. A partir do trabalho de Nussinov e Jacobson [114], o algoritmo passa a incorporar também o cálculo de energia livre nas regras de preenchimento da matriz de adjacência da equação (6.1). Quando Xia et al. [29] disponibilizam um conjunto bastante completo de energias para RNA novos trabalhos surgiram [119–122] em que, dado o conhecimento sobre diversas outras estruturas secundárias, maximizar o número de pares de base deixa de ser prioridade e passa-se a buscar a minimização da energia livre do sistema. Estes trabalhos associam um dado valor de energia de formação para os diversos elementos que ocorrem na estrutura secundária. Por exemplo, caso ocorra um *bulge* de tamanho dois na estrutura secundária, a energia da fita irá ser aumentada em 2.43 kcal/mol. Estes valores experimentais que auxiliam na predição de estruturas foram obtidos através de experimentos de desnaturação de fitas que sabidamente continham tais elementos [123].

6.3.1 Softwares de predição de estrutura secundária

Hoje os softwares mais utilizados na literatura para predição de estruturas secundárias (`mfold` $\approx$ 9200 citações, `RNAfold` $\approx$ 2000 citações) possuem diversos recursos adicionais se comparados à época de sua publicação. Eles contam, por exemplo, com pareamentos do tipo GU, com termos de emparelhamento renovados por experimentos mais precisos feitos posteriormente [123], e com diversos outros softwares auxiliares que permitem modificar as funções básicas dos programas originais (ou seja, programas para obter estruturas subóptimas, como veremos na seção 6.4).

`RNAfold` é um programa pertencente ao pacote `Vienna` [110, 119, 124] que foi implementado como programa isolado e também como *web server*, ou seja com uma interface *web* razão pela qual se tornou rapidamente a ferramenta padrão de biologia molecular. Além das informações sobre energia livre de Gibbs o algoritmo incorpora também vínculos estruturais $\beta(i,j)$ na avaliação das regras de preenchimento

$$E(i,j) = \min \begin{cases} E(i,j-1) \\ \min_{i<k<j} E(i+1,k-1) + E(k+1,j) + \beta(i,j) \end{cases} \tag{6.2}$$

Ou seja, avalia-se se as energias adjacentes além do fator $\beta(i,j)$. Aqui a função do vínculo estrutural

[1] Não há garantia de que essa é a energia mínima real. Ela representa a menor energia obtida para aquela estrutura, a partir dos parâmetros utilizados no modelo

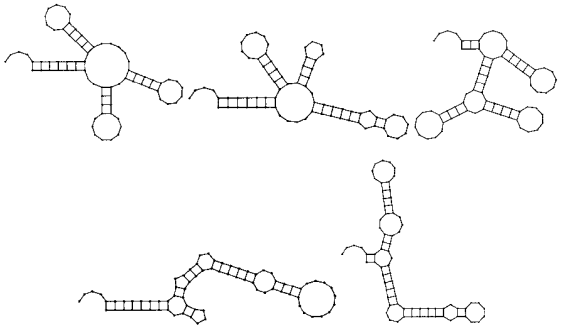


Figura 6.8
As estruturas subóptimas de uma sequência de tRNA de levedura. Figura da referência 125.

$\beta(i,j)$ é dar um peso a informações experimentais sobre determinadas estruturas. Por exemplo, um par i,j pode ter uma energia livre desfavorável, mas sabe-se que por dados experimentais que ela estrutura é muito recorrente, pode-se então alterar o valor $\beta(i,j)$ para refletir este fato. Assim, o algoritmo implementado no `RNAfold` é na verdade um híbrido usando fatores estruturais, quando estes são conhecidos, e energias livres nos demais casos.

6.4 Estruturas secundárias subóptimas

Como vimos na seção 6.3, mesmo usando a regra de preenchimento mais básica, a estrutura secundária pode assumir mais de uma forma para uma mesma estrutura primária. No entanto, mesmo com a estrutura sendo diferente, o número de pares de base formados permanece igual. Em parte isto ocorre quando $E(i,j)$ é usado em incrementos discretos. No entanto, quando se usa a energia livre de Gibbs, ou seja uma variável contínua, em geral se obtém apenas uma estrutura tipo MFE. Mas aqui há outro problema a considerar: para a energia livre de Gibbs também há uma incerteza experimental inerente nos dados que são usados. Por exemplo, para os dados de Xia et al. [29] a incerteza experimental de ΔG é na ordem de 3.2% por par de base. Assim, introduzimos aqui o conceito de estruturas subóptimas, como sendo aquelas próximas da MFE usando por argumento a incerteza experimental. Na figura 6.8 vemos uma sequência de tRNA de uma levedura com alguns dos possíveis formatos de suas estruturas subóptimas. Podemos observar que as estruturas subóptimas, mesmo obtidas de uma mesma sequência primária podem ser completamente diferentes, contendo *loops* e *stems* em número e tamanho muito variados.

Estruturas subóptimas foram apresentadas pela primeira vez no trabalho de Williams e Tinoco [126] onde as regras de preenchimento são ajustadas para encontrar estruturas na vizinhança da MFE. Ou seja, embora a justificativa das estruturas seja a incerteza experimental dos parâmetros, esta não é incluída explicitamente nos cálculos. Ao invés disto considera-se aquelas estruturas na

vizinhança de um termo de energia livre arbitrário $\delta(i,j)$ da energia mínima E_{MFE}. A determinação é feita no algoritmo de *backtrace*, discutido na seção 6.3, alterando a regra

```
if S(i+1,j-1)+E(i,j)=S(i,j):
```

para

```
if abs(S(i+1,j-1)-S(i,j)) <= E(i,j)+delta(i,j)
```

onde `delta(i,j)` corresponde à $\delta(i,j)$ e é o incremento em relação à E_{MFE}. Como agora se abrem mais possibilidade de estruturas, o número de estruturas aumentará exponencialmente em relação ao incremento $\delta(i,j)$ e em relação ao tamanho da sequência. Para sequências muito grandes calcular todas as estruturas subóptimas pode se tornar praticamente impossível.

O *software* para realizar predição de estruturas subóptimas chamado `RNAsubopt`, parte do pacote `Vienna` [110, 119], foi desenvolvido originalmente por Wuchty et al. [125]. Neste caso o algoritmo procura obter todas as estruturas possíveis de uma dada sequência primária percorrendo todas as possibilidades do *backtrace* num intervalo δ acima da energia mínima E_{MFE}. Na tabela 6.2 mostramos um exemplo de uma sequência de 40 nucleotídeos e as estruturas subóptimas calculadas num intervalo de $\delta = 2$ kcal/mol. Aqui podemos observar algumas peculiaridades do cálculo de estruturas secundárias Por exemplo, a estrutura 6 é a de menor energia, -5.80 kcal/mol, mas tem menos pares de base formados que as estruturas 1–5. A razão disto é que as estruturas 1–5 possuem um *loop* adicional `((.......))`. Estes *loops* entram no cálculo da energia livre com um valor positivo bem maior, +5–7 kcal/mol, eliminando o benefício da estabilidade gerada pelos pares adicionais gerados.

O *software* `RNAsubopt` pode também gerar um *ensemble* de estruturas secundárias sorteadas aleatoriamente usando como probabilidade o peso de Boltzmann $p = e^{-E(i,j)/k_bT}$ onde k_b é a constante de Boltzmann e T é a temperatura em 37°C. Ou seja, uma estrutura é gerada sem limite δ, e é retornada ou não com uma probabilidade p pelo programa. Efetuando este sorteio repetidas vezes obtemos um conjunto muito grande de estruturas possíveis como mostrado na tabela 6.3. Note que várias, mas não todas as estruturas da tabela 6.2 ocorrem na tabela 6.3, mostrando claramente que `RNAsubopt` não retorna a totalidade das estruturas possíveis mesmo quando geradas em grande quantidade.

	sequência/estrutura	ΔG (kcal/mol)
1	AUAAACGAGAGUUCAUCGUUUAACGCCCCACACGCUUUCA	
2	.((((((((.......)))))))).((.......)).....	-4.80
3	.((((((((....).)))))))).((.......)).....	-3.80
4	.((((((((.......)))))))).((.......)).....	-4.80
5	.((((((((....).)))))))).((.......)).....	-3.80
6	.((((((((.......))))))))................	-5.80
7	.((((((((....).))))))))................	-4.80
8	..(((((((.......)))))))................	-4.10

Tabela 6.2
Exemplo de estruturas subóptimas calculadas com o *software* `RNAsubopt`, num intervalo $\delta = 2$ kcal/mol da $E_{\text{MFE}} = -5.80$ kcal/mol. Observe que as estruturas 1–4 possuem dois *loops*. As estruturas estão apresentadas na ordem em que são geradas pelo programa.

	sequência/estrutura	ocorrências
	AUAAACGAGAGUUCAUCGUUUAACGCCCCACACGCUUUCA	
6	.(((((((.....))))))).....	4018
1	.(((((((.....)))))))..((.....))....	3541
8	..((((((.....)))))).....	541
4	.(((((((.....)))))))..((.....))....	469
	..((((((.....))))))...((.....))....	402
	...(((((.....))))).....	140
	......((((((....(((...)))).....)))))).	136
	......((((((.....))))))).	123
	.((((((.....)))))).....	100
	.((((((.....))))))..((.....))....	97
	.(((((((.....)))))))(.((.....)).)...	58
	...(((((.....)))))....((.....))....	54
	..((((((.....))))))..((.....))....	34
	......((((((.....((.........))..)))))).	33
	.(((((((.....)))))))..(.....).....	30
	......((((((....((.....))....)))))).	29
	.(((((((.....)))))))(.((.....)))....	26
	.((.((((.....)))).))..((.....))....	25
	....(((.....)))).....((.....))....	23
	..((((((.....))))))((.((.....)).))..	23
	....(((.....)))).....(.((.....)).)....	22
	..((((((.....)))))).....	22
	..((((.....)))).....	21
	......((((((.....(.....).....)))))).	20
	......((((((.....((.....)).....)))))).	20
	...(((((.....)))))...((.....))....	20
	..((((((.....))))))((.((.....))))...	20
	.((((((.....))))))..((.....))....	17
	..((((.....))))...((.....))....	14
	..((((((.....))))))...(.....).....	12
	.(((((((.....)))))))(((.....)).)....	10
	......(((((.....(((.....)))))))).	9
	...(((((.....)))))..((.((.....)).)))..	9
	......((((((.....(.....))))))).	8
	.((..(((.....)))..)).((.....))....	8
	.(((((..((.....))))))).....	8
	..((((((.....))))))..(.....).....	7
	.((.((((.....)))).)).....	7
	(..((((((.....))))))...((.....))....	6
	......((((((.....(.....)..)))))).	4
	......((((((.....((...)).....)))))).	4
	...(((((.....)))).)).....	4
	..((((..(.....))))))...((.....))....	4
	.(((.(((.....)))).))).....	4
	......(((((.....)))))).	3
	.(.(((((.....)))))).)..((.....))....	3
	.....(((.....))).....	2
	.....(((.....)))......((.....))....	2
	...((((.....)))).....	2
	....((((.....))))....((.....))....	1
	...((((.....))))...((.....))....	1
	...(((((.....)))))(((.((.....)).)).)	1
	..(((((.....)))))((.((.....)).)))..	1
	..((((((.....))))))(.((.....)))....	1
	.(((((.....))))).....	1

Tabela 6.3
Exemplo de ocorrências de estruturas subóptimas calculadas com o *software* `RNAsubopt` em 10200 realizações. Os índices à esquerda são as estruturas correspondentes da tabela 6.2.

Código do grupo	L769L	S836S	S904S	rs76759170	rs3026785
TCCGT (WT)	-	-	-	-	-
GCCGT	×	-	-	-	-
TCGGT	-	-	×	-	-
GCCGC	×	-	-	-	×
GTCAC	×	×	-	×	×

Tabela 6.4
Descrição de quais SNPs ocorrem em cada grupo de pacientes [128]. O código também representa a sequência que varia entre os mRNA. É importante notar que estes polimorfismos são SNPs sinônimos, ou seja, não codificam uma proteína diferente da original.

6.4.1 Estruturas subóptimas de mRNA de MTC

Aqui retomamos a questão do mRNA do carcinoma medular da tireoide (*medullary thyroid carcinoma*, MTC) que discutimos brevemente na seção 6.1. O câncer MTC to tipo MEN 2 dos pacientes que foram analisados pelos nossos colaboradores da UFRGS apresentava um tipo especial de polimorfismo, polimorfismo de nucleotídeo único (*single nucleotide polimorfism*, SNP). SNP é um polimorfismo em que uma única base nitrogenada de uma dada região é trocada (i.e. A por C). Polimorfismos são simplesmente formas alternativas que um mesmo gene pode ter na população de uma dada espécie, que represente uma fatia da população que seja superior a 1% (caso contrário, seria chamada mutação) [127]. Mais de um SNP pode ocorrer simultaneamente, em regiões diferentes, no gene rearranjado durante a transfecção (*REarranged during Transfection*, RET) dos pacientes. Isto está indicado na tabela 6.4, em que são apresentadas as SNPs dos pacientes estudados. O tipo mais comum de gene tipo TCCGT chamando aqui de gene selvagem (*wild-type*, WT) .

A identificação da sequência genética de um paciente com câncer MTC MEN2 é essencial para ajudar a identificar familiares que possam vir a desenvolver a doença. Quando um indivíduo carrega os SNPs em seu código genético é altamente provável que algum parente próximo também o tenha. Pode-se identificar quais pessoas são portadoras do SNP, e caso necessário encaminhar para tratamentos ou usar estes dados como auxílio para programas de diagnósticos precoces [129].

Os polimorfismos apresentados (veja tabela 6.4) são chamados de “polimorfismos sinônimos”. Isto significa que mesmo que um único nucleotídeo seja trocado, a proteína gerada no final será a mesma. Isto é possível graças à relação entre códons e aminoácidos, que já fora discutida na seção 5.2 desta tese, ou ao SNP se encontrar em uma região que não será traduzida no ribossomo. Uma vez que os SNPs das amostras são sinônimos, os mecanismos de desenvolvimento da doença não devem estar relacionados com a proteína em si. Então, o que levaria o polimorfismo sinônimo, que

gera exatamente mesma proteína, causar um câncer agressivo? Uma das hipóteses levantadas seria a estabilidade termodinâmica do RNA mensageiro relacionado à estes polimorfismos, pois embora a proteína gerada pelo mRNA seja a mesma, a estrutura secundária pode conter diferenças que a tornem mais estável e produza a proteína em maior quantidade ou por tempos mais prolongados.

A região específica do gene RET onde se encontram os SNPs de nosso interesse é a região $3'$ não-traduzida (*$3'$ untranslated region*, $3'$UTR), que possui 5684 nucleotídeos. Como ela é uma região do gene RET em que não ocorre tradução, estas SNPs serão sinônimas. Porém, elas ainda podem afetar a expressão gênica através de mecanismos de controle e interferência, como discutido no capítulo 5.

Como vimos nas seções anteriores, a estrutura secundária é relevante para a função biológica de uma sequência de RNA. Ela foi obtida para cada uma das sequências do gene RET dos grupos de pacientes (tabela 6.4), para que se pudesse verificar a presença de diferenças estruturais entre elas. Porém, em uma sequência de 5684 pares de base, nossos conjuntos de SNPs que são responsáveis pela substituição de 1,2 ou 4 nucleotídeos não causaram uma deformação estrutural relevante ao longo de toda a estrutura secundária. Na verdade, nossos resultados forneceram estruturas secundárias com diferenças estruturais insignificantes entre todos os grupos de pacientes, exceto para a sequência do grupo GTCAC. Logo, a estrutura secundária MFE de cada sequencia não seria o suficiente para concluir por uma maior estabilidade entre os polimorfismos. A figura 6.9 mostra a estrutura secundária da WT e as demais poderem ser vistas no artigo anexado no apêndice E.

Para gerar as estruturas subóptimas através do software `RNAsubopt` temos duas opções, como discutimos na seção 6.4: a primeira é gerar todos os subóptimos possíveis presentes dentro de um intervalo de energia especificado. A segunda é sortear algumas dentre as possíveis estruturas. O segundo método é de extrema utilidade quando temos uma sequência grande o suficiente para tornar o custo de calcular todas as sequências proibitivo. Tentamos utilizar ambas abordagens, mas, naturalmente, para nossas sequências de tamanho superior a 5000 pares de base, a segunda abordagem se provou mais eficaz.

Primeiro calculamos a MFE para todas as estruturas com o `RNAfold` com parâmetros padrão [110], estas ficam na faixa de -2142 a -2138 kcal/mol, ou seja uma diferença de 4 kcal/mol ou 0.2%. Esta diferença é insuficiente para determinar com qualquer margem de segurança qual seria a estrutura secundária mais estável. Em seguida geramos 8000 estruturas subóptimas com o `RNAsubopt` usando a geração por *ensembles* com peso de Boltzmann como discutido na seção 6.4

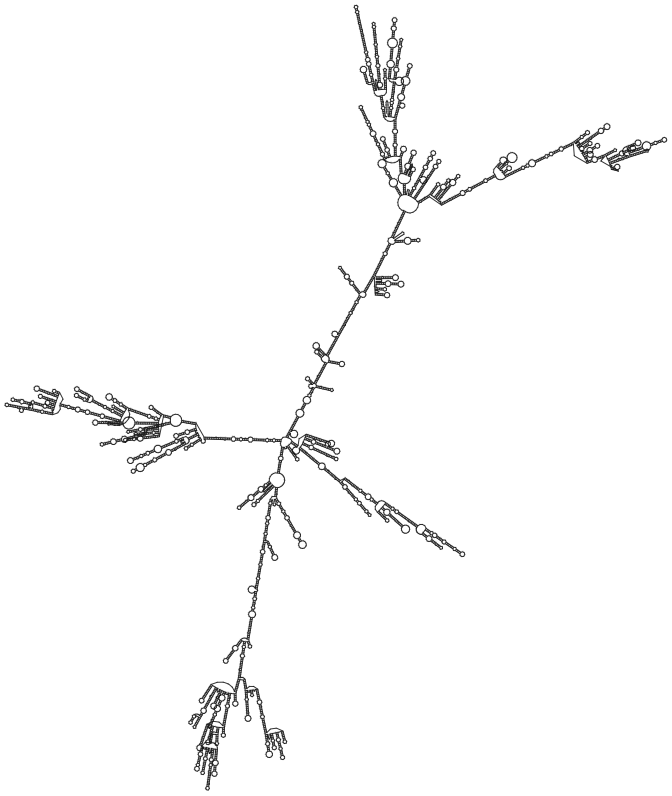


Figura 6.9
Esta é a estrutura secundária de menor energia livre encontrada para a sequência selvagem do proto-oncogene RET. Figura gerada pelo software `RNAfold` com parâmetros padrão [110].

e calculamos a energia livre média $\langle \Delta G_{\text{so}} \rangle$. A figura 6.10 mostra a comparação entre a MFE e $\langle \Delta G_{\text{so}} \rangle$ calculadas desta maneira. A variação de entre as médias das estruturas subóptimas ficou ao redor de 6 kcal/mol. A linha tracejada mostra como deveria ser o comportamento dos valores se todos seguissem o comportamento do gene WT. Note que há um desvio importante desta linha tracejada em especial para a variante GTCAC, esta possui a MFE mais estável mas também os sobóptimos são em média muito mais estáveis do que seria de se esperar se seguisse o gene WT.

Uma pergunta que fica em aberto em relação aos resultados da figura 6.10 é se realmente o cálculo de $\langle \Delta G_{\text{so}} \rangle$ não varia se incluirmos um conjunto diferente de estruturas. Para responder a esta pergunta, nos calculamos $\langle \Delta G_{\text{so}} \rangle$ para dois genes em 8 *ensembles* de 200 estruturas pelo `RNAsubopt`. Quando comparamos o $\langle \Delta G_{\text{so}} \rangle$ dos *ensembles* com a média sobre o total de estruturas, vemos que há uma pequena flutuação geralmente inferior 1 kcal/mol ao redor desta média. Isto mostra que o valor médio dos ensembles flutua, mas sempre se mantém próximo da média obtida.

Uma das razões para se analisar a média dos *ensembles* de estruturas secundárias subóptimas, além da dificuldade em se decidir pela estabilidade termodinâmica de um mRNA somente pela MFE, é que no ambiente celular o mRNA não está completamente livre para forma a estrutura ótima. A ideia é que na célula o mRNA possa formar estruturas intermediárias que seriam similares às estruturas subóptimas [130]. Então a análise representada na figura 6.10 mostra que para o gene

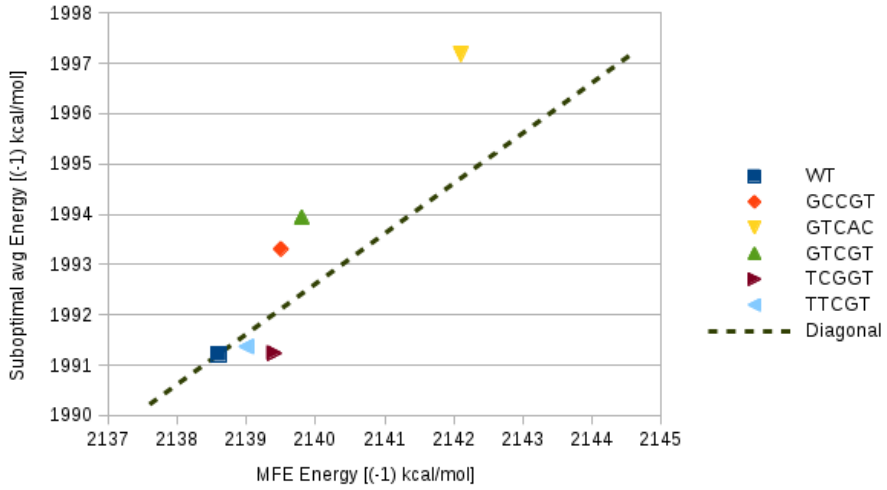


Figura 6.10
Comparação entre a energia média das estruturas subóptimas $\langle \Delta G_{\text{so}} \rangle$ e a MFE. As médias $\langle \Delta G_{\text{so}} \rangle$ foram realizadas sobre 8000 estruturas subóptimas. Note que os eixos estão em unidades de -1 kcal/mol. A linha tracejada seria a tendência se fosse mantida a relação MFE$\times\langle \Delta G_{\text{so}} \rangle$ do gene WT e é colocada aqui apenas como referência.

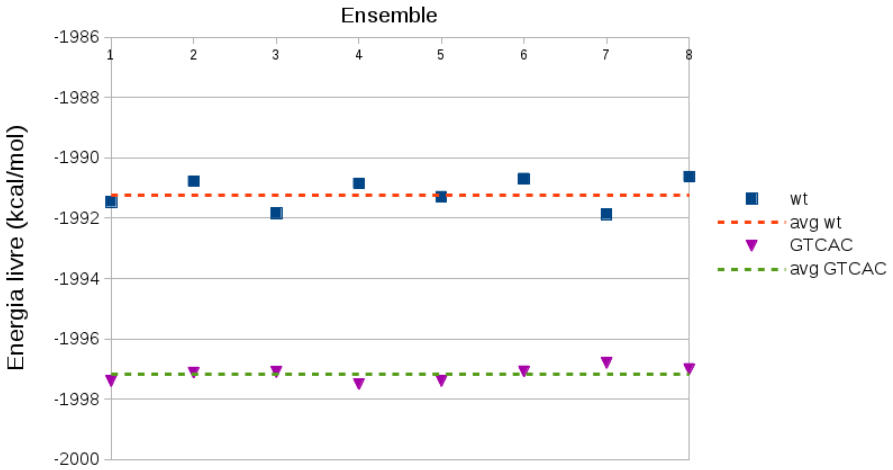


Figura 6.11
Evolução da energia média $\langle \Delta G_{\text{so}} \rangle$ em 8 *ensembles* de 200 estruturas secundárias para o gene WT e CTCAC. As linhas tracejadas são as médias sobre o total das estruturas.

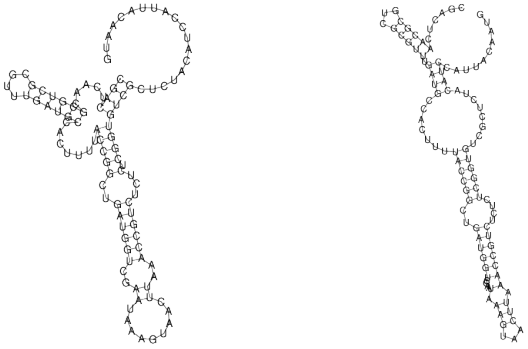


Figura 6.12
Dois exemplos de estruturas secundárias tipo MFEEU obtidas a partir da mesma sequência primária, apenas variando o erro experimental.

GTCAC é mais provável que forme uma estrutura secundária subóptima estável do que qualquer outra variante.

6.5 Influência dos parâmetros experimentais na predição de estruturas secundárias

O trabalho que segue, naturalmente, é verificar se as estruturas subóptimas seriam as mesmas obtidas através da variação dos parâmetros termodinâmicos do modelo de próximos vizinhos (*nearest-neighbour model*, NN). Ou seja, ao invés de gerar estruturas subóptimas permitindo uma variação δ acima da energia livre mínima (*minimum free energy*, MFE), nós vamos calcular a MFE variando os próprios parâmetros de energia livre. Faremos a variação dos parâmetros dentro do erro experimental esperado e compararemos os conjuntos. Nós vamos chamar as estruturas geradas desta maneira de MFE com incerteza experimental (*MFE within experimental uncertainty*, MFEEU). Como motivação, mostramos na figura 6.12 dois testes de MFEEU realizados variando os dados de entrada dentro da faixa dos erros experimentais. Mesmo para uma sequência relativamente curta, a diferença nas estruturas já é muito grande. Assim, é de se esperar que esse efeito seja relativamente maior para sequências grandes, em especial para as sequências de mRNA discutidos na seção anterior. A ideia de variar os parâmetros experimentais não é nova, diversos autores observaram que a incerteza experimental deveria influenciar os resultados de estruturas secundárias [131, 132]. Mas não há ainda um trabalho sistemático explorando este efeito em detalhe, provavelmente porque se conjectura que um conjunto de estruturas tipo MFEEU geradas variando os parâmetros seja equivalente a calcular as estruturas subóptimas mantendo os parâmetros fixos [126]. Um outro argumento levantado por Zuker et al. [133] diz que o erro relativo da energia livre de Gibbs do modelo de próximos vizinhos diminui na medida que a sequência aumenta de tamanho. Considerando que a variação total de

energia livre seja

$$\Delta G_T = \sum_i^N \Delta G_i \tag{6.3}$$

e a propagação de um erro aleatório σ_i do parâmetro ΔG_i, resultaria no erro total

$$\sigma_T = (\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_N^2)^{1/2} \tag{6.4}$$

Se considerarmos uma energia livre uniforme ΔG, um erro constante de 10%, $\sigma = 0.1\Delta G$, para uma sequência de tamanho 100 teremos $\Delta G_T = 100\Delta G$ e $\sigma_T = \Delta G$. Ou seja o erro relativo da energia livre para uma sequência de 100 seria de 1% apenas, no que Zuker et al. [133] conclui:

> *This implies that for large sequences such as those shown [...], it should be sufficient to search suboptimal structures within a few percent of the free energy of the optimal structure.*

À partir desta constatação, a pergunta que nós nos fizemos foi se o número de estruturas secundárias tipo MFEEU aumentaria, ou não, com o tamanho da sequência se introduzíssemos a incerteza experimental na análise. A segunda pergunta foi se o conjunto de estruturas subóptimas seria similar ao conjunto de estruturas MFEEU, já que foi esta a afirmação de Zuker et al. [133].

Um outro motivador para entender melhor a influência dos parâmetros termodinâmicos é o fato de que a comparação entre estruturas secundárias preditas com aquelas determinadas experimentalmente tem uma taxa de acerto baixa. Por exemplo Mathews et al. [123] relata uma taxa ao redor de 65% para o acerto do número correto de pares de base. Este número é ainda menor, 40%, se for contabilizado a estrutura os seus principais elementos, principalmente *loops* e *bulges*, sendo que apenas 3% tem a estrutura completamente predita [134, 135]. Isto leva a seguinte pergunta: seria possível achar um conjunto de parâmetros, dentro a incerteza experimental, que melhoraria estas predições?

6.5.1 Procedimento

Primeiro geramos 30000 novos parâmetros de entrada no programa `RNAfold`, utilizando uma randomização dos dados que garantisse uma distribuição gaussiana para os mesmos. O erro experimental utilizado foi 0.1 kcal/mol para pares de base canônicos [29] e 0.5 kcal/mol para *bulges* [136]. Na implementação, os nossos *scripts* recriam o arquivo de parâmetros do `RNAfold`, exemplificado no apêndice C, ou seja criamos 30000 destes arquivos. Quando garantimos que os dados estão modificados de acordo com o erro experimental, nós podemos recalcular através do `RNAfold` as novas

	sequência/estrutura	ocorrências
	AUAAACGAGAGUUCAUCGUUUAACGCCCCACACGCUUUCA	
6	. (((((((. . . . .))))))))	26779
7	. ((((((((. . . .) .)))))))	2648
4	. (((((((. . . . .)))))))) . ((. . . . .))	399
1	. (((((((. . . . .))))))) . . ((. . . . .))	171
	. . . . . ((((((. . . . (((. . .))) . . . . .)))))) .	2
	. ((((((. . . . .))))))	1

Tabela 6.5
Exemplo de ocorrências de estruturas MFEEU para uma sequência de tamanho 40 nucleotídios. Os índices à esquerda são as estruturas correspondentes da tabela 6.2.

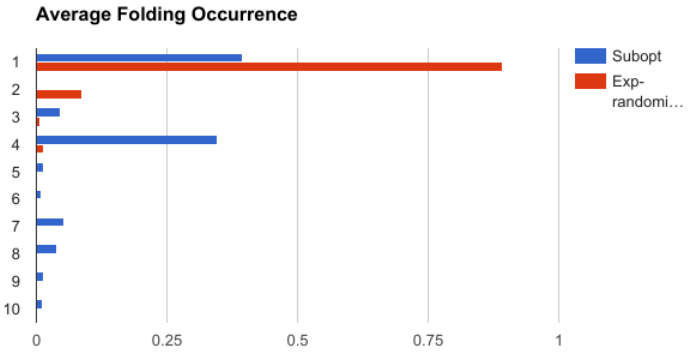


Figura 6.13
Histograma dos índices γ da equação (6.5) para estruturas subóptimas (barras azuis) e tipo MFEEU (barras vermelhas

estruturas, agora do tipo MFEEU. O código computacional que executa a randomização é um *script* na linguagem Perl e é apresentado no apêndice B. Fizemos isto para um conjunto de 30000 novos parâmetros, realizando o empacotamento de 1000 sequências de variados tamanhos. Na tabela 6.5 mostramos a distribuição de ocorrências de estruturas tipo MFEEU da mesma sequência de RNA que usamos na tabela 6.2 das estruturas subóptimas. A comparação das tabelas evidencia que a segunda estrutura mais comum MFEEU da tabela 6.5 nem sequer aparece nas subóptimas tabela 6.2. Já a segunda mais comum das subóptimas (sequência 1) é bastante rara na tabela 6.5. A tabela 6.6 mostra os conjuntos lado a lado onde introduzimos o índice médio estruturas distintas γ

$$\gamma = \frac{\text{número de estruturas distintas geradas}}{\text{total de estruturas geradas}} \tag{6.5}$$

e também a figura 6.13 mostra as diferenças em um histograma. Logo, os subóptimos não podem ser tratados como se fossem estruturas válidas diante o erro experimental. Assim, apenas este exemplo, numa sequência bem pequena mostra que a afirmação de Zuker et al. [133] provavelmente não tem fundamento.

Perceba que das várias estruturas MFEEU feitas temos uma grande maioria igual ao *folding*

Estrutura Secundária	Subopt	Exp-randomização	Índice
AUAAACGAGAGUUCAUCGUUUAACGCCCCACACGCUUUCA	γ(MFE)	γ(MFEEU)	
.(((((((.......))))))).....	0.3939	0.8923	1
.(((((((((....).))))))).....	0	0.0883	2
.(((((((.......)))))))..((.......))....	0.3472	0.0133	3
.(((((((.......)))))))..((.......))....	0.046	0.0057	4
......((((((....(((...)))........)))))).	0.0133	0.00006	5
.((((((.........))))))....	0.0098	0.00003	6
..((((((.......)))))).....	0.053	0	7
..((((((.......))))))...((.......))....	0.0394	0	8
...((((.......)))).....	0.0137	0	9
......((((((.....................)))))).	0.0121	0	10

Tabela 6.6
Tabela comparativa mostrando qual a média de vezes que cada empacotamento aparece no conjunto calculado γ, seja ele um conjunto de subóptimos ou de MFEEU. As sequências foram ordenadas pelo índice de mais provável, para menos provável.

original da estrutura MFE. Computamos então que esta sequência apresentou 6 estruturas MFEEU distintas em 30000 rodadas do nosso algoritmo, ou seja, para esta sequência temos o índice médio de estruturas distintas $\gamma = 2\times10^{-4}$. Na figura 6.14 nós temos o índice médio de estruturas MFEEU distintas para sequências variando de 200 a 800 nucleotídios. As duas curvas presentes se referem à variar os parâmetros dentro do erro experimental completo (*full*), e a outra curva quando variamos apenas com a metade do erro experimental (*half*). Aqui o propósito em considerar metade do incerteza experimental era de estimar o impacto que teria se houvessem dados experimentais melhores. Note que o índice γ aumenta linearmente com o comprimento das sequências, e até 800 nucleotídios não mostra nenhum sinal de alcançar um limite. Infelizmente, o tempo de execução do `RNAfold` aumenta com o cubo do tamanho das sequências, então não foi possível calcular estruturas maiores dentro da infra-estrutura computacional disponível.

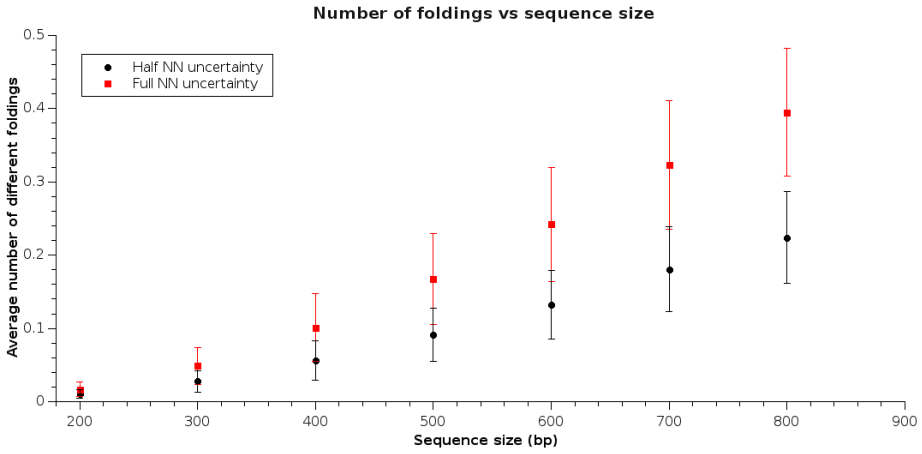


Figura 6.14
Índice de estruturas distintas do tipo MFEEU γ. No eixo vertical temos $\gamma = N/30000$ que é a razão entre a quantidade N de *foldings* MFEEU diferentes pelo valor total de parâmetros testados. Quadrados pretos referem-se às estruturas calculadas com incerteza experimental completo, e pontos vermelhos referem-se a estruturas com metade da incerteza experimental. As barras de erro são os desvios padrão em relação às 30000 estruturas realizadas.

6.5.2 Otimização dos parâmetros termodinâmicos de estruturas secundárias

Como vimos na seção anterior, dentro da incerteza experimental existe uma grande quantidade de estruturas possíveis. E esta quantidade cresce com o tamanho da sequência de RNA. Por outro lado, como havíamos comentado, na comparação entre estruturas secundárias verificadas experimentalmente e aquelas calculadas há uma taxa de acerto muito pequena. Já que as estruturas dependem tão fortemente com os parâmetros experimentais, será que algum conjunto em particular poderia fornecer uma predição melhor? Para poder responder à esta pergunta precisamos primeiro introduzir alguns conceitos que medem a qualidade da predição de estruturas secundárias.

6.5.3 Índices da qualidade de predição

Introduzimos primeiro o conceito de "sensitividade" (*sensitivity*) como sendo

$$\text{Sensitividade} = \frac{\text{número de pares-de-base corretamente preditos}}{\text{número de pares-de-base verificados no experimento}} \tag{6.6}$$

que mostra quão "sensível" os parâmetros são em prever corretamente o número de pares de uma estrutura secundária concheida. O conceito do valor positivo predito (*positive predicted value*, PPV), às vezes também chamado de "especificidade",

$$\text{PPV} = \frac{\text{número de pares-de-base corretamente preditos}}{\text{número de pares-de-base preditos}} \tag{6.7}$$

Aqui mantivemos a nomenclatura usual para este típo de análise [137]. Cada um dos índices auxilia no entendimento de um aspecto da qualidade da predição. A sensitividade nos ajuda a saber qual proporção dos pares-de-base corretamente preditos estão de fato presentes como um par-de-base na estrutura real, é uma medida da semelhança com o experimento. Já a medida de PPV nos indica qual proporção entre o número de pares-de-base preditos que são corretos, para o total (corretos e incorretos), ou seja, ela é uma medida que nos informa mais sobre a própria predição. A predição ideal é aquela que conseguir atingir 100% em ambos índices.

6.5.4 Procedimento de otimização

Além da análise da afirmação de Wuchty et al. [125], sobre a natureza dos subóptimos, ainda testamos qual a influência da variação dos parâmetros na qualidade de predição. Para verificar a qualidade precisávamos de um conjunto de estruturas experimentalmente determinadas. Utilizamos um extenso conjunto de estruturas experimentalmente conhecidas, disponibilizado pelo grupo do Prof David Mathews.

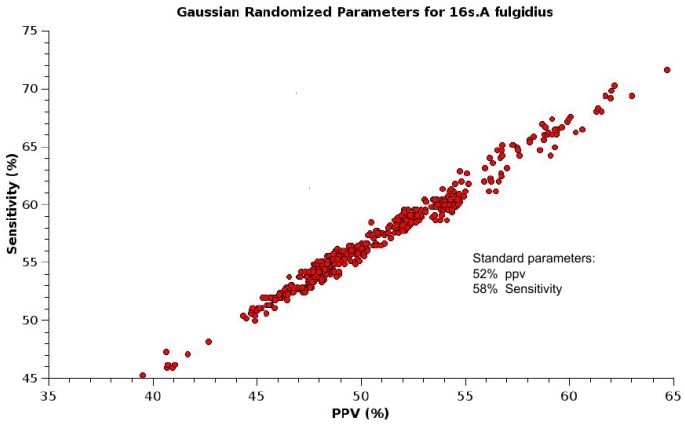


Figura 6.15
Cada ponto do gráfico representa o valor do PPV e sensitividade obtido por aquele conjunto de parâmetros para os genes do domínio 1 da sequência 16S da espécie *A. fulgidius*.

Verificamos, ainda, dentre todos os 30000 conjuntos de parâmetros experimentais fisicamente equivalentes que foram gerados, qual era o conjunto que obtém o maior índice de acertos em termos de sensitividade e PPV. Esta engenharia reversa nos permitiu redefinir os parâmetros padrão utilizados no programa para aumentar o índice de acerto geral das predições. Em outros termos: unir as informações dos experimentos de cristalografia de raio-x e de termodinâmica de desnaturação nos permitiu obter melhores parâmetros para determinação do empacotamento das estruturas secundárias formadas pelos ácidos nucleicos.

Primeiramente, escolhemos uma das estruturas secundárias conhecidas de uma dada sequência real de alguma espécie. Começamos pelos genes do domínio 1 da sequência 16S da espécie *A. fulgidius*. Na figura 6.15 cada ponto representa um conjunto de parâmetros específico, dentre os 30000 possíveis. Tal ponto é indexado pelos valores dos índices de qualidade discutidos na seção 6.5.3. O que obtivemos é que o conjunto padrão disponível não é o que fornecia o empacotamento mais adequado para aquela sequência específica. Nós possuíamos conjuntos de parâmetros que forneciam estruturas secundárias mais parecidas com a real do que os que são utilizados comumente pelos softwares.

O procedimento seguinte foi calcular uma média de qualidade para todos os empacotamentos determinados experimentalmente de que nós dispúnhamos. Cada conjunto de parâmetro randomizado dentro do erro experimental foi utilizado para fazer o empacotamento de todas as sequências, de todas as espécies. O que obtivemos foi um índice de qualidade médio do PPV e da sensitividade. Os resultados dos valores médios de PPV e sensitividade para cada conjunto de parâmetros, em relação à uma média para todas as espécies podem ser vistos na figura 6.16.

O ponto preto no gráfico representa o conjunto padrão de parâmetros termodinâmicos usados no `RNAfold` (SENS-stdpar), ou seja, o valor médio de PPV e sensitividade obtido por tal conjunto

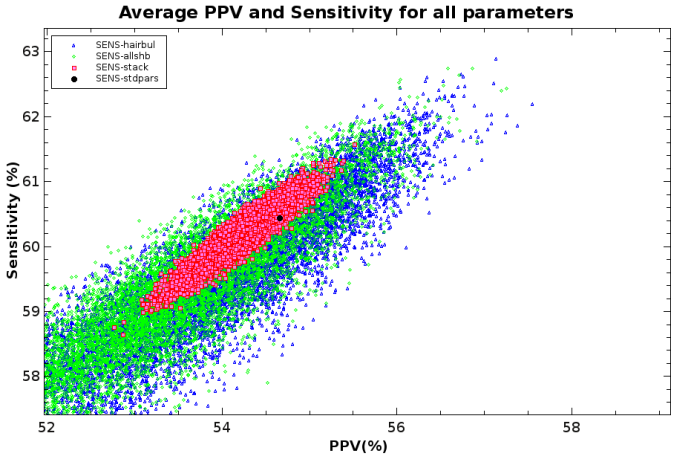


Figura 6.16
Sensitividade em função de PPV. Cada ponto do gráfico representa o valor médio de PPV e sensitividade obtido por aquele conjunto de parâmetros (sendo tal média sobre o conjunto de todas estruturas experimentais disponíveis).

para todas estruturas experimentais disponibilizadas. Os triângulos azuis representam os parâmetros que tiveram apenas os termos de hairpin e bulges randomizados (SENS-Hairbul), os quadrados vermelhos representam os parâmetros que tiveram apenas os termos de empilhamento randomizados (SENS-Stack), e, por fim, os losangos verdes representam os conjuntos de parâmetros em que tanto os parâmetros de empilhamento quanto os de bulges e *hairpin* foram randomizados (SENS-allshb). Note que a variação na qualidade das predições é muito menor quando modificamos os parâmetros de empilhamento do nosso modelo. Isto sugere que tais parâmetros estão bem definidos experimentalmente. Em contrapartida, variar os parâmetros termodinâmicos de formação dos *hairpins* e *bulges* teve uma influência muito maior na qualidade das predições, similarmente a variar todos os parâmetros simultaneamente. Isto nos faz crer que tais parâmetros não estavam tão bem definidos quanto os de empilhamento, e que agora, após nosso processo de engenharia reversa, possuímos valores mais próximos dos ideias para estas estruturas. O melhor conjunto de parâmetros obtidos está descrito no apêndice C.

6.5.5 Investigação da melhora nos índices de qualidade

Após detectarmos que havíamos obtidos um conjunto de parâmetros melhores do que os disponíveis até então, partimos para averiguação de quais são os efeitos destes parâmetros no conjunto de dados experimentais. As hipóteses levantadas eram de que ele melhoraria os piores resultados, aprimoraria todos sutilmente, ou melhoraria os resultados de algumas sequências específicas apenas. O que descobrimos é que os resultados com baixo índice de acerto, no caso, sequências cujo empacotamento predito diferia demasiadamente do evidenciado experimentalmente, que acarretava um baixo valor de PPV e sensitividade, praticamente ficaram inalteradas. Isto pode ser visto na

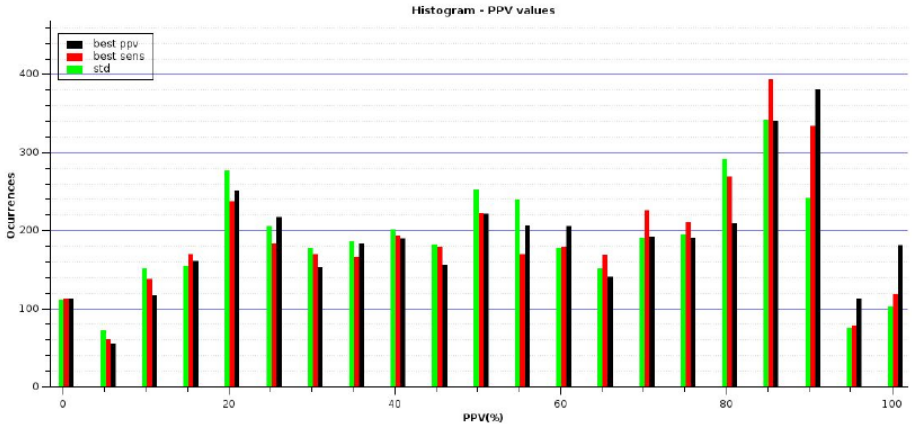


Figura 6.17
O comportamento da quantidade de sequências com diferentes índices de qualidade está evidenciada nos gráficos acima. As barras verde, vermelha e pretas indicam diferentes parâmetros que foram utilizados para obtenção destas distribuições. A coluna *std* se refere aos parâmetros padrão, a coluna *best-sens* se refere ao conjunto de parâmetros que apresentou o melhor índice médio de sensitividade, e a coluna *best-ppv*, analogamente, se refere ao que apresentou melhor índice médio de PPV.

parte esquerda da figura 6.17. O conjunto de parâmetros termodinâmicos com melhores resultados de PPV, como pode ser visto em tal figura, possui a mesma quantidade de estruturas preditas grosseiramente do que o conjunto de parâmetros padrão (marcados como std no gráfico). Por outro lado, uma dada quantidade de sequências que apresentavam índices de acertos bons passou a possuir índices de acerto melhores ainda. Devido à isso vemos na figura 6.17 um aumento da quantidade de sequências dispostas na região direita do gráfico, com altíssimos índices de confiabilidade — alto valor de PPV.

6.6 Conclusão

Mostramos neste capítulo o papel importante da estrutura secundária na determinação da estabilidade de RNA. Em particular, mostramos que o cálculo de estruturas subóptimas pode auxiliar na determinação da estabilidade de grandes sequências mRNA relacionadas a oncogenes. Mostramos que a distribuição de estruturas secundárias subóptimas não corresponde ao conjunto de estruturas quando se varia a incerteza experimental, ao contrário do que tem sido afirmado na literatura [133]. O número de sequências distintas tipo MFEEU cresce linearmente com o tamanho da sequência e até 800 nucleotídios não mostrou sinal de chegar num nível de saturação. Finalmente, mostramos que é possível otimizar o valor dos parâmetros termodinâmicos para obter um conjunto que produza previsões melhores, mas não está claro se este conjunto possui significado físico.

Conclusões

As conclusões desta tese são divididas em duas seções: cada uma delas referentes à uma parte do trabalho realizado. Primeiramente, discutiremos as conclusões referentes ao trabalho com a inosina, capítulo 4. Este trabalho, que deu início aos meus estudos no campo de biofísica de ácidos nucleicos, nos permitiu determinar valores para as interações da inosina em pareamento com cada uma das bases do DNA, e consigo mesma. Os resultados são animadores, tendo em vista que estão de acordo com outras técnicas teóricas, e diversas observações experimentais. Encontramos que a inosina se pareia com as bases canônicas do DNA formando duas ligações de hidrogênio com cada uma delas. As ligações provavelmente possuem geometrias diferentes, dado nossos resultados para a interação de empilhamento entre bases distintas da fita. As nossas evidências sugerem que uma das bases distorce a cadeia de açúcar-fosfato, tornando as interações de empilhamento diferentes do usual esperado para as bases nitrogenadas quando pareadas. Evidenciamos ainda que este fato pode tornar a inosina uma base útil para a atividade e desenvolvimento de sequências de DNA atuando como retificadores térmicos.

Em segundo lugar, temos as conclusões sobre a análise dos parâmetros termodinâmicos usados para predição de estruturas secundárias de RNA, e sua relação com as estruturas subótimas, capítulo 6. Este trabalho foi iniciado por uma discussão advinda de outro trabalho do qual participei neste período de doutoramento sobre a estrutura do RNA de pacientes de câncer de tireoide — tal trabalho resultou um artigo presente no apêndice E desta tese. Fizemos uma análise estatística para averiguar a afirmação de que "As estruturas subótimas representam as possibilidades de empacotamento possíveis de ocorrer a uma sequência ao considerarmos o erro experimental das medidas dos parâmetros termodinâmicos". O que verificamos é que o conjunto das estruturas subótimas é bem distinto do conjunto obtido por variação direta do erro experimental MFE com incerteza experimental (*MFE within experimental uncertainty*, MFEEU). O conjunto subótimo apresentava muito mais variabilidade de empacotamentos possíveis do que o outro conjunto, com proporções estatísticas

diferentes dele. Como os subótimos não são, necessariamente, uma alternativa para o empacotamento das sequências, então nosso estudos devem buscar a melhora das predições usuais de outras maneiras. Testamos, assim, a hipótese de que talvez os conjuntos de parâmetros termodinâmicos usados como entrada nos softwares de predição pudessem ser melhorados. Encontramos o melhor conjunto de dados por engenharia reversa utilizando as estruturas secundárias reais obtidas por raios X e NMR. Este conjunto está disponibilizado no apêndice C.

Apêndice A

Minha trajetória profissional durante o doutorado

Este pequeno esboço de *Curriculum Vitae* entra como parte desta tese para que eu pudesse relatar diversas experiências de aprendizado que geralmente não são diretamente vistas em artigos ou similares. Espero que seja um estímulo para futuros alunos de pós-graduação e graduação para que invistam em suas carreiras, e em participar ativamente das comunidades científicas às quais pertencem.

Comecei meu doutrado em outra área, e após um ano migrei para a biofísica computacional. O primeiro trabalho com o qual me envolvi foi a determinação dos parâmetros do modelo Peyrard-Bishop para inosina. Durante o primeiro semestre me dediquei à esta tarefa e participei do meu primeiro encontro científico na área: uma escola internacional de modelagem computacional de sistemas biológicos em Porto Alegre, em 2013. Esta escola foi extremamente produtiva, onde pude conhecer diversos pesquisadores de outras áreas e suas linhas de pesquisa. Nesta mesma viagem tive contato com o grupo de estudos do câncer MCT-MEN2A na UFRGS, liderado pela professora Ana Luiza Maia. Este encontro foi frutífero para ambas partes e culminou, eventualmente, em um artigo científico feito em colaboração com a equipe da UFRGS.

Em 2014 e 2015 também participei do Encontro Nacional de Física da Matéria Condensada, apresentando meus avanços nas pesquisas que o nosso grupo estava desenvolvendo à época. Foram experiências extremamente satisfatórias e enriquecedoras. O Encontro Nacional é o evento científico brasileiro que mais reúne físicos em nosso país. Poder participar dele, em cada uma das vezes, foi um grande privilégio.

Participei de escolas de introdução à dinâmica molecular de sistemas biológicos em Belo Horizonte, oferecida pelo departamento de biologia da UFMG. Nele aprendi conceitos rudimentares da área de dinâmica molecular. O mini-curso fora ministrado por professor convidado da Unibersidade Federal de Brasília, que nos deu diversos *insights* sobre esta área.

O II Workshop on Biomolecular-Theory and Experiment Interplay foi outro diferencial na minha formação. Nos reunímos com diversos pesquisadores que modelavam sistemas biológicos em escala molecular utilizando as mais diversas técnicas - da mecânica estatística, à mecânica quântica através de um misto de DFT e Dinâmica Molecular. Além das análises teóricas, as análises dos experimentos que geram os dados que modelamos também pôde ser discutida. Este encontro permitiu a interação de físicos, químicos, matemáticos, biológos, em um contexto muito produtivo.

Finalmente, em 2017, participei do último congresso do meu período de doutoramento: o X-meeting em Belo Horizonte. Este congresso internacional trouxe pesquisadores de diversos países, estudando diversos temas com a genética e a computação como ponto em comum. Neste evento eu pude ver pesquisas de altíssimo nível que estavam sendo desenvolvidas no Brasil, Índia, Estados Unidos, etc. Após esta data, a confecção da tese se tornou prioridade até o fim do período de doutorado, defendido no início de 2018.

Apêndice B

Códigos Perl utilizados

Neste apêndice deixo registrado parte dos scripts de programação em linguagem Perl utilizados ao longo deste trabalho. Eles não serão discutidos em detalhes, mas podem servir como exemplos para futuros estudantes fazerem seus próprios scripts.

B.1 Average-repetitions-per-size.pl

Este script tem como finalidade verificar quantas vezes um mesmo resultado de um cálculo de empacotamento aparece dentre as 30000 tentativas feitas para uma dada sequência. Ele também percorre todos os tamanhos de sequência disponíveis na pasta de input.

```
1  #!/usr/bin/perl -w
2
3  my @filename_list = `ls ./analysis/*counter*`;
4  #print @filename_list;
5
6  my %details_diff_per_file_per_size = ();
7  my %total_sum_value_per_size = ();
8  my %total_of_structures_counted = ();
9
10 foreach my $filename (@filename_list)
11   {
12   #print "$filename ";
13   $filename =~ m/sequence-([0-9]*)-([0-9]*)-counter.txt/;
14   my ($index,$size) =  ($1,$2);
15   #print "Estes sao os identificadores number: $index, size: $size\n";
16
17   my @diff_content = `tail -n +4 $filename`;
18   #print "@diff_content \n";
19   my $n_diff_structures = scalar @diff_content;
20   #print "Existem $n_diff_structures estruturas diferentes aqui\n\n";
21   #print "$content[1]\n";
22
23   $total_sum_value_per_size{$size} += $n_diff_structures;
24   $total_of_structures_counted{$size} += 1;
25   push @{ $details_diff_per_file_per_size{$size} },
26     "$n_diff_structures (arquivo $index) \n";
27  }
28
29 my %averages = ();
30 my %std_devs = ();
31
32 foreach my $size (keys %details_diff_per_file_per_size)
```

```
33    {
34    my $square_sum = 0;
35    my $std_dev = 0;
36    my $avg = sprintf("%.2f", $total_sum_value_per_size{$size}
37                              /$total_of_structures_counted{$size});
38
39
40      foreach my $item ( @{ $details_diff_per_file_per_size{$size} } )
41        {
42        $item  =~ m/([0-9]*) \(/;
43        my $item_value = $1;
44        $square_sum += ($item_value - $avg)**2;
45        }
46
47    $std_dev = sprintf("%.2f", sqrt ($square_sum/$total_of_structures_counted{$size}));
48
49    $averages{$size} = $avg;
50    $std_devs{$size} = $std_dev;
51    }
52
53  foreach $size (keys %details_diff_per_file_per_size)
54  {
55  open (SAVE_FILE, ">./analysis/files-$size-avg-std-dev.txt") or die "falhou tudo";
56  print SAVE_FILE "#media desvio tamanho\n";
57  print SAVE_FILE "$averages{$size} $std_devs{$size} $size \n----------
58  \n----------\n----------\n";
59  print SAVE_FILE "Para tamanho $size, todos os arquivos detalhados sao:\n
60  @{ $details_diff_per_file_per_size{$size} }\n";
61  close SAVE_FILE;
62  }
63
64  sub remove_energy_of_bracket_structure()
65  {
66  my ($string) = @_;
67
68  $string =~ s/([\-]|\s)([0-9]+)\.([0-9]+)//g;
69  $string =~ s/(\s\(\s?\))//g;
70
71  return $string;
72  }
73
74  #
75  sub obtain_sizes()
76  {
77  my @size_list = ();
78  my $i = "020";
79
80    while ($i < 820)
81    {
82    push @size_list, "$i";
83    $i+=20;
84    $i = sprintf("%03d",$i);
85
86    }
87
88  return @size_list;
89  }
```

B.2 repeated-structures.pl

Este é o programa que diferenciava, para um dado tamanho, quantas estruturas eram únicas.

```
#!/usr/bin/perl -w

my @filename_list = `ls ./analysis/*together*`;
print @filename_list;

foreach my $filename (@filename_list)
  {
  $filename =~ m/sequence-([0-9]*-[0-9]*)-together.fa/;
  my $name = $1;

  if (-e "./analysis/sequence-$name-counter.txt")
    {print "./analysis/sequence-$name-counter.txt áj existe \n"; next;}

  print "Este é o nome novo: $name\n";

  my %hash_sequences =();

  my @content = `cat $filename`;

  my $content_size = scalar @content;

    for (my $i=2; $i <= $content_size; $i+=3)
      {
      chomp $content[$i]; #brackets
      my $check = remove_energy_of_bracket_structure($content[$i]);
      #print "$check \n";

      if (exists $hash_sequences{$check}) {$hash_sequences{$check} += 1;}
      else {$hash_sequences{$check}=1;}
      }

  open (SAVE_FILE, ">./analysis/sequence-$name-counter.txt") or die "falhou tudo";
  print SAVE_FILE "$filename\n$content[1]";

    foreach my $key (keys %hash_sequences)
      {
        print SAVE_FILE "$key: repeated $hash_sequences{$key} times \n";
      }
  close SAVE_FILE;
  }
```

B.3 separating-results.pl

Em várias etapas dos procedimentos era necessário agrupar ou separar as diversas sequências que compunham um dado grupo de RNAs para entrada de informações nos softwares disponíveis nos pacotes do RNAfold. Esta rotina separava com um loop duplo os conteúdos de alguns dos resultados.

```
#!/usr/bin/perl -w

my @filenames = create_valid_filenames_list();

foreach my $filename (@filenames)
{
my ($size,$type,$index) = obtain_name_details($filename);

if (-e "./analysis/sequence-1000-$size-together.fa")
  {print "Arquivos do tipo (...)-$size-together.fa áj existem \n"; next;}

my @file_content = `zcat $filename`;

```

```
14 my ($i,$j) = (1,1);
15
16 foreach $line (@file_content)
17   {
18     $j = sprintf("%03d",$j);
19     my $save_file = "./analysis/sequence-$j-$size-together.fa";
20     open (SAVE_FILE, '>>', "$save_file") or die "Couldn't open file $save_file";
21     if ($i==1) {chomp $line; $line .= "-within-file-$type-$index \n";}
22     print SAVE_FILE "$line";
23     $i++;
24     if ($i > 3) {$i=1; $j++; close SAVE_FILE;}
25   }
26 }
```

B.4 calculador-ppv-sensitividade.pl

Como o nome da seção já diz, era esta parte do script que permitia o cálculo dos nossos índices de qualidade. Neste exemplo de código, vemos a utilização de diversas subrotinas. Elas deixam o código mais "limpo e automaticamente comentado", como sugerem as melhores práticas de programação.

```
1  #!/usr/bin/perl
2
3  system("rm ./together-archivell/*");
4
5  @fasta_list = `ls ./results-archivell/*2???.txt.gz`;
6
7  foreach $fasta_item (@fasta_list)
8  {
9    @fasta_content = `zcat ./results-archivell/$fasta_item`;
10   separador_many_structures_into_singles($fasta_item,@fasta_content);
11 }
12
13 @mestres_list = `ls ./together-archivell/*.ct`;
14
15 print "lista dos together\n", @mestres_list, "\n\n";
16
17 foreach $mestre_item (@mestres_list)
18 {
19   print "analisando termo $mestre_item\n";
20   calculador_ppv_item($mestre_item);
21   @lista_final_pronta = criar_lista_final_ppv_sens_nome($mestre_item);
22   $mestre_item =~ /together-archivell\/(.*)-together\.ct/; #obter o nome
23
24   open(ARCHIVE,">./archivell-final-ppv-listings/$1-final-ppv-listings.txt");
25   print ARCHIVE @lista_final_pronta;
26   close ARCHIVE;
27
28   system ("gzip -f ./archivell-final-ppv-listings/$1-final-ppv-listings.txt");
29
30   gzippar_arquivo_together_ja_lido($1);
31 }
32
33 sub separador_many_structures_into_singles()
34 {
35 ($fasta_item,@array) = @_;
36 chomp $fasta_item;
37 $fasta_origin = substr ($fasta_item, 0, -7);
38
39 while (@array) #quando o array for nulo, termina ele
40   {
```

```
41    $fileheader = shift @array; #shift translada o array e o primeiro termo, $array[0], é
    retornado.
42    $sequence = shift @array;
43    $structure = shift @array;
44
45    $filename = encontrar_nome_arquivo_correto($fileheader);
46
47    #print "Agora irei salvar o nome que encontrei: $filename\n";
48
49    open(ARCHIVE,">./individual-bra-archivell/$filename-within-$fasta_origin.bra");
50    print ARCHIVE "$fileheader","$sequence","$structure";
51    close ARCHIVE;
52
53    criar_ct_do_arquivo_e_compactar
54  ("./individual-bra-archivell/$filename-within-$fasta_origin.bra");
55
56    adicionar_ct_no_arquivo_mestre
57  ("./individual-ct-archivell/$filename-within-$fasta_origin.ct");
58
59    compactar_bra_individual_e_deletar_original
60  ("./individual-bra-archivell/$filename-within-$fasta_origin.bra");
61    }
62  }
63
64
65  sub encontrar_nome_arquivo_correto()
66  {
67  ($string) = @_;
68  $nome_limpo = limpar_nome_fasta($string);
69
70  $grep_result = `grep -H '$nome_limpo' ./experimentais-dot-brackets-archivell/*.bra`;
71  #fornece resultado nome_arquivo_encontrado:trecho_buscado
72
73  $grep_result =~ /experimentais-dot-brackets-archivell\/(.*):/;
74
75  $match = substr ($1, 0, -4);
76  # o únmero 1 corresponde ao match da regular expression. O -4 remove o .bra do fim.
77
78  return $match;
79  }
80
81  sub calculador_ppv_item()
82  {
83  ($string) = @_; #mestre_item ./together-archivell/$name-together.ct
84  chomp $string;
85  $string =~ /\/together-archivell\/(.*)-together.ct/;
86  $name = $1;
87
88  system ("./scorer-ppv-sens $string ./experimentais-ct-archivell/$name.ct
89  ./together-archivell/$name-together-ppv-scores.scr");
90  return 1;
91  }
92
93  sub limpar_nome_fasta()
94  {
95  ($string) = @_;
96  $string =~ />(\s*)(.*)/;
97  $match = $2;
98  return $match;
99  }
100
101 sub criar_ct_do_arquivo_e_compactar()
102 {
```

```
103 ($string) = @_;
104 $ctname = $string;
105 $ctname =~ s/bra/ct/g;
106
107 system ("./dot2ct $string $ctname");
108 system ("gzip -f $ctname");
109 }
110
111 sub adicionar_ct_no_arquivo_mestre()
112 {
113 ($string) = @_; #./individual-ct-archivell/$filename-within-$fasta_origin.ct
114
115 $string =~ /\/individual-ct-archivell\/(.*)-within-(.*).ct/;
116 $nome = $1;
117 $origem = $2;
118 system ("zcat $string >> ./together-archivell/$nome-together.ct" );
119 system ("echo $nome-within-$origem >> ./together-archivell/referencia-$nome-together.txt");
120 return 1;
121 }
122
123 sub compactar_bra_individual_e_deletar_original()
124 {
125 ($string) = @_;
126 system("cat $string | gzip > $string.gz");
127 system("rm $string");
128 return 1;
129 }
130
131 sub criar_lista_final_ppv_sens_nome()
132 {
133 ($string) = @_;
134 chomp $string;
135 $string =~ /\/together-archivell\/(.*)-together.ct/;
136 $name = $1;
137
138 @lista_origens = `cat ./together-archivell/referencia-$name-together.txt`;
139 @lista_valores = `cat ./together-archivell/$name-together-ppv-scores.scr`;
140 @tabela_final = ();
141
142 for ($j =0; $j < scalar @lista_origens; $j++)
143 {
144   if ($j == 0) {$tabela_final[0] = "#item sensitivity ppv name-of-origin-structure\n";}
145
146   shift @lista_valores; shift @lista_valores; shift @lista_valores;
147   #joga fora 3 linhas indesejadas
148
149   $temp = shift @lista_valores;
150   $temp =~ /= (.*)%/;
151   $ppv_value = $1;
152
153   $temp = shift @lista_valores;
154   $temp =~ /= (.*)%/;
155   $sens_value = $1;
156
157   shift @lista_valores; #joga a última linha, em branco, fora
158
159   $tabela_final[$j] .= "$j $sens_value $ppv_value $lista_origens[$j]";
160 }
161 return @tabela_final;
162 }
163
164 sub gzippar_arquivo_together_ja_lido
165 {
```

```
166 ($name_to_zip) = @_;
167
168 @list = `ls -I *gz ./together-archivell/*$name_to_zip*`;
169 foreach $item (@list)
170 {
171 system ("gzip -f $item");
172 }
173 return;
174 }
```

B.5 ppv-sensitivity.pl

Para calcular os índices de qualidade da grande massa de arquivos que geramos tivemos que fazer diversas adaptações, como por exemplo salvar arrays perl em disco. Mesmo que cada processo individual rodasse rapidamente, gravar eles em disco consumia um tempo desmasiadamente longo. O que fizemos foi rodar vários deles em pacotes, e gravar tais aglomerados em disco sem converter para texto comum - gravando diretamente um estado suspenso da execução dos scripts. Estes cálculos foram divididos em três etapas, cada uma rodando por mais de um dia, quando enfim conseguimos a análise final. A seguir, mostro os scripts das três etapas.

B.5.1 ppv-sensitivity-step1.pl

```
1  #!/usr/bin/perl -w
2  use strict;
3
4  system("rm ./together-archivell/*");
5
6  my @fasta_list = `find ./results-archivell/ `; # AQUI ESTA O TESTE
7  shift @fasta_list;
8
9  print "lista dos fastas\n", @fasta_list, "\n\n";
10
11 foreach my $fasta_item (@fasta_list)
12 {
13   print "zcat $fasta_item";
14   my @fasta_content = `zcat $fasta_item`;
15
16   separador_many_structures_into_singles($fasta_item,@fasta_content);
17
18 }
19
20 sub separador_many_structures_into_singles()
21 {
22 my ($fasta_item,@array) = @_;
23
24 chomp $fasta_item;
25
26 $fasta_item =~ /\.\/results-archivell\/(.*).txt.gz/;
27 my $fasta_origin_name = $1;
28
29 print $fasta_origin_name;
30
31 while (@array) #quando o array for nulo, termina ele
32   {
33   print "\nRodando o íincio do while\n\n";
34
35   my $fileheader = shift @array;
36   #shift translada o array e o primeiro termo, $array[0], é retornado.
37   my $sequence = shift @array;
```

```
38    my $structure = shift @array;
39
40    ($structure, $fileheader) =
41  remover_energia_do_folding_e_colocar_no_header($structure,$fileheader);
42
43    my $filename = encontrar_nome_arquivo_correto_pela_seq($sequence);
44
45    open(ARCHIVE,">./individual-bra-archivell/$filename-within-$fasta_origin_name.bra");
46    print ARCHIVE "$fileheader","$sequence","$structure";
47    close ARCHIVE;
48
49    compactar_bra_individual_e_deletar_original
50  ("./individual-bra-archivell/$filename-within-$fasta_origin_name.bra");
51    }
52
53  }
54
55  sub encontrar_nome_arquivo_correto_pela_seq()
56  {
57  my ($string) = @_;
58  chomp $string;
59
60  my $grep_result = `grep -H '^$string\$' ./experimentais-dot-brackets-archivell/*.bra`;
61  #fornece resultado nome_arquivo_encontrado:trecho_buscado
62
63  $grep_result =~ /experimentais-dot-brackets-archivell\/(.*)\.bra:/;
64
65  #$match = substr ($1, 0, -4);
66  # o numero 1 corresponde ao match da regular expression. O -4 remove o .bra do fim.
67  my $match = $1;
68
69  return $match;
70  }
71
72  sub remover_energia_do_folding_e_colocar_no_header()
73  {
74  my ($structure,$fileheader) = @_;
75   $structure =~ s/\s\(([\-])([0-9]+)\.([0-9]+)\)\s//g;
76
77  my $energy_value = " ---- Folding energy: ($1$2.$3)\n";
78  chomp $fileheader;
79  $fileheader = $fileheader . $energy_value;
80  $structure = $structure . "\n";
81  return ($structure,$fileheader);
82  }
83
84  sub limpar_nome_fasta()
85  {
86  my ($string) = @_;
87  $string =~ />(\s*)(.*)/;
88  my $match = $2;
89  return $match;
90  }
91
92  sub compactar_bra_individual_e_deletar_original()
93  {
94  my ($string) = @_;
95  system("cat $string | gzip > $string.gz");
96  system("rm $string");
97  return 1;
98  }
```

B.5.2 ppv-sensitivity-step2.pl

```
 #!/usr/bin/perl -w
use strict;
use warnings;
use Storable;

#Choose a path, comment the other
my $path = "/home/rodolfomaxi"; #for sbatch
#my $path = "."; #for zion, dalek command line

sub compose_ct_name_to_be_saved
{
my ($ctname) = @_;
$ctname =~ s/\.bra$/.ct/g;
$ctname =~ s/individual-bra-archivell/individual-ct-archivell/g;

return $ctname;
}

sub save_ppv_sheet_on_disk
{
my ($sheetname,$ref_ppv_sheet) = @_;
$sheetname =~ s/\.bra$/.dat/g;
$sheetname =~ s/individual-bra-archivell/individual-sheet-archivell/g;
open (OUTSHEET, ">$sheetname") or die "\nCouldn't open $sheetname";
print OUTSHEET "i p s name  -----  i=index p=ppv s=sensitivity
name=(foldfile-with-experimentalfile)\n";
print OUTSHEET @{$ref_ppv_sheet};
close OUTSHEET;

return;
}

sub criar_ct_temp_do_arquivo
{
my ($string) = @_;
my $ctname = $string;
chomp $string;
$ctname =~ s/\.bra$/.ct/g;
my $cmd="./dot2ct $string $ctname";
system ($cmd);
return;
}

sub construction_final_ct_array
{
my ($bra_file,$full_value) = @_;
$full_value =~ /([0-9])([0-9])([0-9])([0-9])/;
my $value = "$1";

my @final_ct_array;

my $array_complete_file_ref = retrieve("$bra_file");
#Estrutura do arquivo, do né-simo elemento:
# $array_complete_file_ref[n][0] = header com energia
# $array_complete_file_ref[n][1] = sequence
# $array_complete_file_ref[n][2] = structure
# $array_complete_file_ref[n][3] = nome

for (my $j=0;$j < scalar @{$array_complete_file_ref};$j++)
  {
  open (OUT, ">$path/tempfolder/set$value/temp$j-$full_value.bra");
  print OUT $array_complete_file_ref->[$j][0] if (exists $array_complete_file_ref->[$j]);
  print OUT $array_complete_file_ref->[$j][1] if (exists $array_complete_file_ref->[$j]);
```

```
64    print OUT $array_complete_file_ref->[$j][2] if (exists $array_complete_file_ref->[$j]);
65    close OUT;
66    criar_ct_temp_do_arquivo("$path/tempfolder/set$value/temp$j-$full_value.bra");
67
68    #Estrutura do arquivo, do né-simo elemento:
69    # $final_ct_array[n][0] = nome do arquivo de folding alternativo
70    # $final_ct_array[n][1] = úcontedo ct do folding
71    # $final_ct_array[n][2] = nome da estrutura experimental ao qual ele deve ser comparado
72    # $final_ct_array[n][3] = comando para calcular o scorer
73
74    #0
75    $final_ct_array[$j][0] = $array_complete_file_ref->[$j][3];
76    $final_ct_array[$j][0] =~ s/\.bra$/.ct/g;
77
78    #1
79    $final_ct_array[$j][1] = "unknown";
80    $final_ct_array[$j][1] = `cat $path/tempfolder/set$value/temp$j-$full_value.ct`;
81
82    #2
83    $final_ct_array[$j][0] =~ /(.*)-within-(.*)/;
84    $final_ct_array[$j][2] = "$1.ct";
85
86    #3
87    if (-e "$path/experimentais-ct-archivell/$final_ct_array[$j][2]")
88      {
89      $final_ct_array[$j][3] =
90  "./scorer-ppv-sens $path/tempfolder/set$value/temp$j-$full_value.ct
91  $path/experimentais-ct-archivell
92  /$final_ct_array[$j][2] $path/tempfolder/set$value/temp-$full_value-ppv-value.dat";
93      }
94    else
95      {
96      $final_ct_array[$j][3] = "echo Erro - nao existia o arquivo experimental
97  $path/experimentais-ct-archivell/$final_ct_array[$j][2] >
98  $path/tempfolder/set$value/temp-$full_value-ppv-value.dat";
99      }
100   }
101   return @final_ct_array;
102 }
103
104 sub construction_final_ppv_sheet
105 {
106 my ($ref_final_ct_array,$full_value) = @_;
107 $full_value =~ /([0-9])([0-9])([0-9])([0-9])/;
108 my $value = "$1";
109
110 my @final_ppv_sheet;
111
112 for (my $j=0;$j < scalar @{$ref_final_ct_array};$j++)
113   {
114   system ("$ref_final_ct_array->[$j][3]");
115
116   my @info_collector = `cat $path/tempfolder/set$value/temp-$full_value-ppv-value.dat`;
117   if ($info_collector[0] =~ /Erro - ãno existia o arquivo experimental/)
118     {
119     $final_ppv_sheet[$j] = "$j 0 0 Erro experimental com -----
120  $ref_final_ct_array->[$j][0]-with-$ref_final_ct_array->[$j][2] ----- ]]\n";
121     print $final_ppv_sheet[$j];
122     next;
123     }
124
125   #Sensitivity line: Sensitivity: 21 / 21 = 100.00%
126   my $sens_value;
```

```
127   if ($info_collector[3] =~ /\=(\s*)([0-9]*)\.([0-9]*)\%/)
128 {$sens_value = "$2" . "." . "$3" . "%";}
129   else {$sens_value = "0.0%";}
130
131   #PPV line: PPV:         21 / 23 = 91.30%
132   my $ppv_value;
133   if ($info_collector[4] =~ /\=(\s*)([0-9]*)\.([0-9]*)\%/)
134 {$ppv_value = "$2" . "." . "$3" . "%";}
135   else {$ppv_value = "0.0%";}
136
137   my $names;
138   if ($sens_value eq '0.0%' || $ppv_value eq '0.0%')
139 {$names = "----- $ref_final_ct_array->[$j][0]-with-$ref_final_ct_array->[$j][2]
140 ----- WARNING\n";}
141   else
142 {$names = $names . "----- $ref_final_ct_array->[$j][0]-with-$ref_final_ct_array->[$j][2]
143 -----\n";}
144
145   #Store values in sheet
146   $final_ppv_sheet[$j] = "$j $ppv_value $sens_value $names";
147   }
148   return @final_ppv_sheet;
149 }
150
151 my ($fasta_origin_name) = @ARGV or
152 die "Need argument, single number (refers to set_ folder number)\n";
153
154 $fasta_origin_name =~ /([0-9])([0-9])([0-9])([0-9])/;
155 my $value = "$1";
156 my $full_value = $fasta_origin_name;
157
158 mkdir "$path/individual-ct-archive11/set$value/"
159 unless -d "$path/individual-ct-archive11/set$value/" ;
160 mkdir "$path/individual-sheet-archive11/set$value/"
161 unless -d "$path/individual-sheet-archive11/set$value/" ;
162 mkdir "$path/tempfolder/set$value/" unless -d "$path/tempfolder/set$value/";
163
164 #Collect bra_files
165 my @bra_files;
166 @bra_files = `find $path/individual-bra-archive11/set$value/*$fasta_origin_name.bra `;
167
168 unless (@bra_files) {print "No files in set$value folder\n"; die;}
169
170  #Transform files
171 foreach my $bra_file (@bra_files)
172 {
173 my $ctname = compose_ct_name_to_be_saved($bra_file);
174
175 print " starting $bra_file\n";
176 print " will save on $ctname";
177
178 my @final_ct_array = construction_final_ct_array($bra_file,$full_value);
179 my @final_ppv_sheet = construction_final_ppv_sheet(\@final_ct_array,$full_value);
180
181 #Save everything in disk
182 store \@final_ct_array, "$ctname";
183 system ("gzip -f $ctname");
184
185 save_ppv_sheet_on_disk($bra_file,\@final_ppv_sheet);
186
187 print "saved $bra_file";
188
189 system ("rm $path/tempfolder/set$value/temp*-$full_value.ct");
```

```
system ("rm $path/tempfolder/set$value/temp*-$full_value.bra");
system ("rm $path/tempfolder/set$value/temp-$full_value-ppv-value.dat");
}
```

B.5.3 ppv-sensitivity-step3.pl

```
 #!/usr/bin/perl -w
use strict;
use warnings;
use Storable;

#Choose a path, comment the other
#my $path = "/home/rodolfomaxi"; #for sbatch
my $path = "."; #for zion, dalek command line

my ($value) = @ARGV or die "Need argument, single number (refers to set_ folder number)\n";

my $ref_averages_sheet;

#Collect bra_files
my @bra_files;
@bra_files = `find $path/individual-sheet-archive11/set$value/*.dat |sort`;

unless (@bra_files) {print "No files in set$value folder\n"; die;}

#Extract data and take mean
foreach my $bra_file (@bra_files)
{
chomp $bra_file;

print "messing with $bra_file\n";

my $cmd_ppv = "tail -n +2  $bra_file |awk '{ total += \$2 } END { print total/NR }'";
my $cmd_sens = "tail -n +2  $bra_file |awk '{ total += \$3 } END { print total/NR }'";

print "calculating average of ppv with command:
$cmd_ppv\ncalculating average of sens with command: $cmd_sens\nResults are: \n";

my $ppv_value = `$cmd_ppv`;
chomp $ppv_value;
$ppv_value = sprintf ("%.2f",$ppv_value);

my $sens_value = `$cmd_sens`;
chomp $sens_value;
$sens_value = sprintf ("%.2f",$sens_value);

print "PPV average: $ppv_value, SENS average: $sens_value\n";

my $string_to_save = "$ppv_value $sens_value $bra_file\n";
$ref_averages_sheet -> {$bra_file} = $string_to_save;
print "What will be saved in the final averages sheet,
regarding this file is:$string_to_save";
}

open (AVG_SHEET, ">$path/individual-sheet-archive11/average-ppv-sens-of-set$value.dat");
print AVG_SHEET "PPV SENS FILENAME\n";
foreach my $hashvalue (sort values %{$ref_averages_sheet}) {print AVG_SHEET $hashvalue;}
close AVG_SHEET;
```

B.6 randomize-parameters.pl

Nesta seção mostro como foram randomizados os parâmetros para corresponderem à uma situação de variações dentre de uma margem de erros experimentais (distribuição gaussiana).

```
 #!/usr/bin/perl
use strict;

for (my $i = 1; $i <=10000; $i++)
{
my $index = sprintf("%05s", $i);
print "nome-$index \n";
system ("perl all-together-randomize-gaussian-parameters.pl
> ./parameters/all-s-h-b-$index.par");
system ("perl hairpin-bulge-randomize-gaussian-parameters.pl
> ./parameters/hair-bul-$index.par");
system ("perl stack-parameters-randomize-gaussian.pl
> ./parameters/stack-$index.par");

}
```

B.6.1 all-together-randomize-gaussian-parameters.pl

Os outros códigos são idênticos à este, porém com omissão de quais parâmetros não seriam randomizados.

```
#!/usr/bin/perl
use strict;

my $max_exp_stack_error = 10; #10 corresponds to 0.1 kcal/mol ,
which is the error as discussed in NDDB turner04, xia98
my $max_exp_error_hairpin_bulges = 50; #50 corresponds to 0.5 kcal/mol ,
which is the error as discussed in NDDB turner04, xia98
my $sigma = 1.0;

my @start_stack_params = define_starting_stack_parameters();
my @final_stack_params = randomize_stack_parameters(@start_stack_params);

my @start_hairpin_params  = define_starting_hairpin_parameters();
my @final_hairpin_params = randomize_hairpin_bulge_parameters(@start_hairpin_params);

my @start_bulge_params  = define_starting_bulge_parameters();
my @final_bulge_params = randomize_hairpin_bulge_parameters(@start_bulge_params);

print "## RNAfold parameter file v2.0 \n";
print "#stack \n @final_stack_params \n";
print "#hairpin \n @final_hairpin_params \n";
print "#bulge \n @final_bulge_params \n";

exit;
sub define_starting_stack_parameters
{

# ## RNAfold parameter file v2.0 [these are in units of 10^(-2) kcal/mol]
#
# # stack
#   -240  -330  -210  -140  -210  -210  -140    /* CG */
#   -330  -340  -250  -150  -220  -240  -150    /* GC */
#   -210  -250   130  -050  -140  -130   130    /* GU */
#   -140  -150  -050   030  -060  -100   030    /* UG */
#   -210  -220  -140  -060  -110  -090  -060    /* AU */
#   -210  -240  -130  -100  -090  -130  -090    /* UA */
#   -140  -150   130   030  -060  -090   130    /* NS */

# # stack
#     CG    GC    GU    UG    AU    UA    NS
#   -240  -330  -210  -140  -210  -210  -140    /* CG */
#   -330  -340  -250  -150  -220  -240  -150    /* GC */
```

```
43 #   -210  -250   130  -050  -140  -130   130    /* GU */
44 #   -140  -150  -050   030  -060  -100   030    /* UG */
45 #   -210  -220  -140  -060  -110  -090  -060    /* AU */
46 #   -210  -240  -130  -100  -090  -130  -090    /* UA */
47 #   -140  -150   130   030  -060  -090   130    /* NS */
48
49 my $CG_CG = '-240';
50 my $CG_GC = '-330';
51 my $CG_GU = '-210';
52 my $CG_UG = '-140';
53 my $CG_AU = '-210';
54 my $CG_UA = '-210';
55 my $CG_NS = '-140';
56
57 my $GC_CG = '-330';
58 my $GC_GC = '-340';
59 my $GC_GU = '-250';
60 my $GC_UG = '-150';
61 my $GC_AU = '-220';
62 my $GC_UA = '-240';
63 my $GC_NS = '-150';
64
65 my $GU_CG = '-210';
66 my $GU_GC = '-250';
67 my $GU_GU = ' 130';
68 my $GU_UG = '-050';
69 my $GU_AU = '-140';
70 my $GU_UA = '-130';
71 my $GU_NS = ' 130';
72
73 my $UG_CG = '-140';
74 my $UG_GC = '-150';
75 my $UG_GU = '-050';
76 my $UG_UG = ' 030';
77 my $UG_AU = '-060';
78 my $UG_UA = '-100';
79 my $UG_NS = ' 030';
80
81 my $AU_CG = '-210';
82 my $AU_GC = '-220';
83 my $AU_GU = '-140';
84 my $AU_UG = '-060';
85 my $AU_AU = '-110';
86 my $AU_UA = '-090';
87 my $AU_NS = '-060';
88
89 my $UA_CG = '-210';
90 my $UA_GC = '-240';
91 my $UA_GU = '-130';
92 my $UA_UG = '-100';
93 my $UA_AU = '-090';
94 my $UA_UA = '-130';
95 my $UA_NS = '-090';
96
97 my $NS_CG = '-140';
98 my $NS_GC = '-150';
99 my $NS_GU = ' 130';
100 my $NS_UG = ' 030';
101 my $NS_AU = '-060';
102 my $NS_UA = '-090';
103 my $NS_NS = ' 130';
104
105 my $CGname = "  /* CG */ \n";
```

```
106 my $GCname = "  /* GC */ \n";
107 my $GUname = "  /* GU */ \n";
108 my $UGname = "  /* UG */ \n";
109 my $AUname = "  /* AU */ \n";
110 my $UAname = "  /* UA */ \n";
111 my $NSname = "  /* NS */ \n";
112
113 my @lineCG = ($CG_CG, $GC_CG, $GU_CG, $UG_CG, $AU_CG, $UA_CG, $NS_CG, $CGname);
114 my @lineGC = ($CG_GC, $GC_GC, $GU_GC, $UG_GC, $AU_GC, $UA_GC, $NS_GC, $GCname);
115 my @lineGU = ($CG_GU, $GC_GU, $GU_GU, $UG_GU, $AU_GU, $UA_GU, $NS_GU, $GUname);
116 my @lineUG = ($CG_UG, $GC_UG, $GU_UG, $UG_UG, $AU_UG, $UA_UG, $NS_UG, $UGname);
117 my @lineAU = ($CG_AU, $GC_AU, $GU_AU, $UG_AU, $AU_AU, $UA_AU, $NS_AU, $AUname);
118 my @lineUA = ($CG_UA, $GC_UA, $GU_UA, $UG_UA, $AU_UA, $UA_UA, $NS_UA, $UAname);
119 my @lineNS = ($CG_NS, $GC_NS, $GU_NS, $UG_NS, $AU_NS, $UA_NS, $NS_NS, $NSname);
120
121 my @parameter_display = (@lineCG, @lineGC, @lineGU, @lineUG, @lineAU, @lineUA, @lineNS);
122
123 return @parameter_display;
    #col 0  #col 1  #col 2  #col 3  #col 4  #col 5  #col 6
124
125 }
126 sub define_starting_hairpin_parameters
127 {
128
129 # ## RNAfold parameter file v2.0 [these are in units of 10^(-2) kcal/mol]
130 #
131 #    # hairpin
132 #    INF   INF   INF   540   560   570   540   600   550   640
133 #    650   660   670   680   690   690   700   710   710   720
134 #    720   730   730   740   740   750   750   750   760   760
135 #    770
136
137 my @hairpin1 = ('INF  ','INF  ','INF  ',540,560,570,540,600,550,640,"\n");
138 my @hairpin2 = (650,660,670,680,690,690,700,710,710,720,"\n");
139 my @hairpin3 = (720,730,730,740,740,750,750,750,760,760,"\n");
140 my @hairpin4 = (770, "\n");
141 my @parameter_display = (@hairpin1, @hairpin2, @hairpin3, @hairpin4);
142
143 return @parameter_display;
144 }
145 sub define_starting_bulge_parameters
146 {
147 # ## RNAfold parameter file v2.0 [these are in units of 10^(-2) kcal/mol]
148 #
149 #    # bulge
150 #    INF   380   280   320   360   400   440   460   470   480
151 #    490   500   510   520   530   540   540   550   550   560
152 #    570   570   580   580   580   590   590   600   600   600
153 #    610
154
155 my @bulge1 = ('INF  ',380,280,320,360,400,440,460,470,480, "\n");
156 my @bulge2 = (490  ,500,510,520,530,540,540,550,550,560, "\n");
157 my @bulge3 = (570  ,570,580,580,580,590,590,600,600,600, "\n");
158 my @bulge4 = (610, "\n");
159
160 my @parameter_display = (@bulge1, @bulge2, @bulge3, @bulge4);
161
162 return @parameter_display;
163 }
164 sub randomize_stack_parameters
165 {
166 my @received = @_;
167
```

```
168 foreach my $param (@received)
169 {
170   if ($param =~ /(\/\*)/) {next;} #Busca os finais de linha
171   my ($newparam, $lixo) = randomize_two_parameters($param,0,$max_exp_stack_error);
172   $newparam = sprintf ("%.3d", $newparam);
173   unless ($newparam =~ /\-/) {$newparam = ' ' . $newparam};
174   $param = $newparam . ' ';
175 }
176
177 @received = symmetrize_the_stack_parameters(@received);
178 return @received;
179 }
180
181 sub symmetrize_the_stack_parameters
182 {
183 #the stack parameter is a 6x6 matrix (a array, where each term is also an array).
184 #Where the terms should be symmetric with the diagonal
185
186 my @notsymmetric = @_;
187
188 my @pamat; #acronymn for parameter matrix;
189 my ($i,$j) = (0,0);
190
191 foreach my $nonsym (@notsymmetric)
192 {
193   $pamat[$i][$j] =  $nonsym;
194   if ($j eq 7) {$j=0; $i++; next;}
195   else {$j++; next;}
196 }
197
198  for (my $i=0; $i < 7; $i++)
199    {
200    for (my $j=0; $j < 7; $j++)
201      {
202      if ($i == $j) {next;}
203      else
204         {
205         ($pamat[$i][$j],$pamat[$j][$i]) =
206         overwrite_both_with_greater_abs($pamat[$i][$j],$pamat[$j][$i]);
207         }
208      }
209    }
210
211  my @symmetric0 = ($pamat[0][0],$pamat[0][1],$pamat[0][2],$pamat[0][3],
212 $pamat[0][4],$pamat[0][5], $pamat[0][6],$pamat[0][7]);
213  my @symmetric1 = ($pamat[1][0],$pamat[1][1],$pamat[1][2],$pamat[1][3],
214 $pamat[1][4],$pamat[1][5], $pamat[1][6],$pamat[1][7]);
215  my @symmetric2 = ($pamat[2][0],$pamat[2][1],$pamat[2][2],$pamat[2][3],
216 $pamat[2][4],$pamat[2][5], $pamat[2][6],$pamat[2][7]);
217  my @symmetric3 = ($pamat[3][0],$pamat[3][1],$pamat[3][2],$pamat[3][3],
218 $pamat[3][4],$pamat[3][5], $pamat[3][6],$pamat[3][7]);
219  my @symmetric4 = ($pamat[4][0],$pamat[4][1],$pamat[4][2],$pamat[4][3],
220 $pamat[4][4],$pamat[4][5], $pamat[4][6],$pamat[4][7]);
221  my @symmetric5 = ($pamat[5][0],$pamat[5][1],$pamat[5][2],$pamat[5][3],
222 $pamat[5][4],$pamat[5][5], $pamat[5][6],$pamat[5][7]);
223  my @symmetric6 = ($pamat[6][0],$pamat[6][1],$pamat[6][2],$pamat[6][3],
224 $pamat[6][4],$pamat[6][5], $pamat[6][6],$pamat[6][7]);
225
226  my @symmetric = (@symmetric0, @symmetric1, @symmetric2, @symmetric3,
227 @symmetric4, @symmetric5, @symmetric6);
228
229  return @symmetric;
230 }
```

```
231
232 sub overwrite_both_with_greater_abs
233 {
234  my ($newparam1,$newparam2) = @_;
235  my $greater = 0;
236  if  (abs $newparam1 > abs $newparam2) {$greater = $newparam1;}
237  else {$greater = $newparam2;}
238  return ($greater,$greater);
239 }
240
241 sub randomize_hairpin_bulge_parameters
242 {
243 my @received = @_;
244 foreach my $param (@received)
245 {
246   if ($param =~ /\n/ | $param =~ /INF/ ) {next;} #Busca os finais de linha, ou infinito
247   my ($newparam, $lixo) = randomize_two_parameters($param,0,$max_exp_error_hairpin_bulges);
248   $newparam = sprintf ("%.3d", $newparam);
249   unless ($newparam =~ /\-/) {$newparam = $newparam . '  '};
250   $param = $newparam;
251 }
252 return @received;
253 }
254 sub randomize_two_parameters
255 {
256 my ($x1,$x2,$rand) = @_;
257 my ($rand1,$rand2) = gauss_ran_num($rand);
258
259 my $x1f = $x1 + $rand1;
260 my $x2f = $x2 + $rand2;
261
262 return ($x1f,$x2f);
263 }
264 sub gauss_ran_num
265 {
266 my ($fwhm) = @_;
267 my $two_pi = 6.28318530717958647692528676;
268 my $rand_max = 1;
269
270 my $u1 = rand(1)/1;
271 if ($u1 < 1e-100) {$u1 = 1e-100;}
272 #print "u1 é $u1";
273
274 my $u2 = rand(1)/1;
275 if ($u2 < 1e-100) {$u2 = 1e-100;}
276 #print " u2 é $u2 \n";
277
278 my $R = sqrt((-2) * log($u1));
279 my $theta = $two_pi * $u2;
280
281 my @gaussian_num = ($fwhm * $R * cos($theta), $fwhm * $R * sin($theta));
282 return @gaussian_num;
283 }
284 sub test_if_gaussian
285 {
286   for (my $i=1; $i <= 10000; $i++)
287   {
288     my $j = 130;
289     my $k = 200;
290     my ($newj, $newk) = randomize_two_parameters($j,$k,$max_exp_stack_error);
291     $newj = sprintf ("%.3d", $newj);
292     $newk = sprintf ("%.3d", $newk);
293     print "$i $newj $newk \n";
```

```
294     }
295     exit;
296     return 1;
297 }
```

Apêndice C

Melhores Parâmetros termodinâmicos

Neste apêndice mostro o formato do melhor conjunto de parâmetros que obtivemos para os softwares de empacotamento que utilizam o modelo de próximos vizinhos e energia livre de formação.

O primeiro conjunto é relativo a pares canônicos AU, CG e GU. NS significa *not specified*, para casos como este exemplo: `AACG(NS)T`. Como se trata de próximos vizinhos a interpretação é a seguinte: o parâmetro se refere ao par da linha seguido pelo par da coluna. Por exemplo -335 (segunda linha, primeira coluna) é a energia livre de CG quando seguido de GC. Note que a matriz é simétrica por uma questão de formato do software. A unidade dos parâmetros é 10^{-2} kcal/mol, ou seja um parâmetro -215 quer dizer -2.15 kcal/mol. Isto foi feito para que todos os parâmetros fossem números inteiros o que acelera a execução do software e reduz o espaço em memória.

```
## RNAfold parameter file v2.0
#stack
 -215  -335  -210  -137  -237  -234  -145    /* CG */
 -335  -338  -260  -152  -223  -241  -157    /* GC */
 -210  -260   130  -058  -150  -141   139    /* GU */
 -137  -152  -058   040  -084  -106   041    /* UG */
 -237  -223  -150  -084  -120  -094  -073    /* AU */
 -234  -241  -141  -106  -094  -141  -094    /* UA */
 -145  -157   139   041  -073  -094   134    /* NS */
```

Para estruturas tipo *loop* e *bulges*, os parâmetros são dados por tamanho crescente iniciando por 1. Neste caso, *loops* de tamanho 1, 2 ou 3 são atribuídos o valor infinito tornando-se impossível a sua formação. Aqui também a unidade é 10^{-2} kcal/mol. Note que todos os valores são positivos, ou seja desestabilizam a estrutura secundária.

```
#hairpin
 INF   INF   INF   592   624   530   567   676   507   550
 655   667   741   682   718   709   727   780   734   761
 753   748   702   736   787   768   788   756   793   848
 728

#bulge
 INF   307   243   422   439   424   393   436   503   496
 478   530   539   533   526   456   672   457   593   650
 609   538   726   643   505   569   686   665   657   624
 664
```

Para fins de referência, seguem os parâmetros originais do pacote Viena [119].

```
## RNAfold parameter file v2.0 [these are in units of 10^(-2) kcal/mol]
#stack
 -240  -330  -210  -140  -210  -210  -140    /* CG */
 -330  -340  -250  -150  -220  -240  -150    /* GC */
 -210  -250   130  -050  -140  -130   130    /* GU */
 -140  -150  -050   030  -060  -100   030    /* UG */
 -210  -220  -140  -060  -110  -090  -060    /* AU */
 -210  -240  -130  -100  -090  -130  -090    /* UA */
 -140  -150   130   030  -060  -090   130    /* NS */

#hairpin
 INF   INF   INF   540   560   570   540   600   550   640
 650   660   670   680   690   690   700   710   710   720
 720   730   730   740   740   750   750   750   760   760
 770

#bulge
 INF   380   280   320   360   400   440   460   470   480
 490   500   510   520   530   540   540   550   550   560
 570   570   580   580   580   590   590   600   600   600
 610
```

Apêndice D

Artigo publicado inosina em DNA pelo modelo Peyrard-Bishop

Chemical Physics Letters 631–632 (2015) 87–91

Contents lists available at ScienceDirect

Chemical Physics Letters

journal homepage: www.elsevier.com/locate/cplett

Deoxyinosine mismatch parameters calculated with a mesoscopic model result in uniform hydrogen bonding and strongly variable stacking interactions

Rodolfo Vieira Maximiano[a,b], Gerald Weber[a,*]

[a] Departamento de Física, Universidade Federal de Minas Gerais, 31270-901 Belo Horizonte, MG, Brazil
[b] Centro Federal de Educação Tecnológica de Minas Gerais, 32210-160 Contagem, MG, Brazil

ARTICLE INFO

Article history:
Received 20 March 2015
In final form 20 April 2015
Available online 12 May 2015

ABSTRACT

Here we apply a technique which uses mesoscopic models to calculate hydrogen bond strengths and stacking interactions from experimental deoxyinosine melting temperatures. Our results confirm an uniform hydrogen bond strength for all deoxyinosine mismatches of the same order of AT base pairs, with the only exception of II mismatches which show a reduced hydrogen bonding. The resulting stacking interactions however are far from uniform, with IG showing the largest variations. The highly asymmetric stacking directly influences the duplex stability and calculated opening profiles show that for some mismatches there can be a localised breakdown of the melting cooperativity.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Inosine occurs naturally in tRNA where it pairs equally to cytosine, adenine and uracil. When used in DNA it has found practical applications as a 'universal' base. For instance, in 1985 the use of inosine was already extensively discussed for probe design [1–3], and applications of inosine for PCR primer design were reported [4,5]. More recent examples are its use in sequencing by ligation, a next-generation sequencing technique where the use of inosine is thought to increase the read lengths [6,7] and for directional cloning [8].

Early thermodynamic studies of deoxyinosine were performed by Kawase et al. [9], and Case-Green and Southern [10]. However, most of our current knowledge of the basic properties of deoxyinosine come from comprehensive UV measurements by Watkins et al. [11]. From nearest-neighbour free energy models they derived a generalised stability trend IC > IA > IT ≈ IG > II which is widely used in practical applications [12,8]. Further detailed properties of hydrogen bonds were obtained from NMR [13–15] and X-ray measurements [16,17]. All these studies suggest a configuration of two hydrogen bonds for all inosine mismatches.

From a theoretical point of view only hypoxantine, the nucleobase of inosine, was studied with density functional theory (DFT) [18–20]. However, these results are difficult to correlate with inosine, especially for the lack of the phosphate backbone in those calculations [18–20]. Molecular dynamics were performed by Krepl et al. [21] for one DNA sequence containing inosine IC mismatch who concluded that the reduced stability is not solely due to the reduced number of hydrogen bonds when compared to guanine.

The thermodynamic analysis of oligonucleotides by mesoscopic models employing experimental melting temperatures is emerging as a way to independently study the hydrogen bond and stacking interactions [22,23]. Given the practical interest of inosine it is worthwhile to extend the investigation of its thermal properties with mesoscopic models.

The basis of our mesoscopic model is the Peyrard–Bishop (PB) model [24]. This is a statistical mechanics model where the main features are based on Morse potentials for the hydrogen bonds interaction and stacking interactions between adjacent base pairs along the DNA strand. This model has the advantage to allow us to obtain the hydrogen bonds and stacking interaction strengths from melting temperatures, as we showed for DNA [22] and RNA [23]. Therefore, the PB model overcomes some of the limitations attributed to free-energy models [21,25,26]. Recent results [27] indicate that there is a also a possibility of obtaining structural parameters such as the helicoidal rise step of DNA.

The PB model is under active development and routinely applied to a number of different physical situations involving oligonucleotides. Some recent examples are its use to study mechanical unzipping [28], charge transport in DNA [29,30], the multifractal

* Corresponding author.
E-mail address: gweberbh@gmail.com (G. Weber).

http://dx.doi.org/10.1016/j.cplett.2015.04.045
0009-2614/© 2015 Elsevier B.V. All rights reserved.

properties of DNA denaturation [31], DNA over-stretching transition [32], the behaviour of DNA in external fields [33], and thermal switching in DNA devices [34]. Another interesting example of application is its use in the study of promoters in cyanobacteria [35] and promoter clustering [36]. One interesting property of the model is that it is easily adaptable to describe further interactions such as the influence of solvent [37–39], and is also being used for in depth studies of mathematical properties of this type of systems [40]. Another mesoscopic model of interest is the rodlike polyelectrolyte model which was recently used for the description of salt-dependent DNA melting [41].

Here we analyse the melting temperatures of sequences containing inosine mismatches from Ref. [11] with the Peyrard–Bishop (PB) model. We were able to obtain a complete description of inosine hydrogen bonds and related stacking parameters. The calculated Morse potentials are consistent with a chemical structure formed by two hydrogen bonds, of similar strength as for AT, for all inosine mismatches. In contrast, the calculated stacking interactions were found to be very dissimilar to that of canonical DNA, with the exception of IC mismatches. Further calculations of helix opening profiles show that the strong variations in stacking interaction result in some large opening amplitudes. In the case of IG and IA mismatches, this disturbance can be strong enough to promote a localised break down of the melting cooperativity.

2. Materials and methods

2.1. Melting temperature dataset

The sequences used for the calculations in this work are UV absorption measurements from Watkins et al. [11], composed of 84 sequences in addition to 13 sequences from other sources [2,42]. These sequences were designed in such a way as to cover every possible neighbouring interaction for a given inosine mismatch pair. The experimental error reported for the melting temperature estimations is 0.3 °C for the 84 oligonucleotide duplexes, which have lengths of either 9 or 12 base pairs (bp) [11]. All sequences are shown in supplementary Tables S1–S5. They are presented with their experimental melting temperature, the calculated melting index, and the predicted melting temperature obtained using the newly calculated parameters. For the minimisation the dataset was divided into groups of sequences that contained a particular IX mismatched pair, where X = A, C, G, T or I. This results in smaller datasets specific to each inosine mismatch (one for IA, another for IC, etc). Tandem inosine mismatches were not included in this analysis as they were too few of them for applying the parameter fitting method [22]. Also, the sequences described as NN outliers of Ref. [11] decreased the quality of the results as melting temperature prediction became poorer for all sequences (data not shown), and therefore these sequences were not included in our dataset.

2.2. The Peyrard–Bishop model

We use a PB model [24] with simple harmonic stacking interactions. The main components of the PB model are the hydrogen bond between the bases of a pair, represented by a Morse potential, and the stacking interaction between nearest neighbours, represented by a harmonic potential in this case. For the example of an IA base pair, the corresponding representation of these terms are

$$V(y_{\mathrm{IA}}) = D_{\mathrm{IA}}(e^{-y_{\mathrm{IA}}/\lambda_{\mathrm{IA}}} - 1)^2 \tag{1}$$

which corresponds to a Morse potential between inosine (I) on one strand with the opposing adenine (A) on the other strand, and y represents the relative distance between the bases. The Morse parameter D largely accounts for the hydrogen bond strength, while λ represents the potential width.

Expanding the previous example for an IA base pair adjacent to CG (IApCG), the stacking interaction is given by

$$w(y_{\mathrm{IA}}, y_{\mathrm{CG}}) = \frac{k_{\mathrm{IApCG}}}{2}(y_{\mathrm{IA}}^2 - 2y_{\mathrm{IA}}y_{\mathrm{CG}}\cos\theta + y_{\mathrm{CG}}^2) \tag{2}$$

The elastic constant k represents the interaction between stacked bases, and θ is a small angle (0.01 rad) introduced to prevent numerical divergence in the integration procedure [43].

The average distance for the m-th base pair can then be calculated from

$$\langle y_m \rangle = \frac{1}{Z_y}\int \prod_n y_m e^{-\beta[w(y_n, y_{n-1}) + V(y_n)]} dy_n \tag{3}$$

where Z_y is the classical partition function for this Hamiltonian [44]. Both the partition function Z_y as well as the distance $\langle y_m \rangle$ are calculated numerically with the transfer integral technique [43].

Eqs. (1) and (2) are the potentials used for the calculation of the classical partition function from which we calculate a melting index τ for each sequence [45]. The melting index is calculated from a numerical expansion of the partition function Z_y in terms of the non-diagonal components of the transfer integral matrix. The basis functions of the expansion are taken from a homogeneous sequence composed of CG base pairs only, see Fig. 2 of Ref. [45]. In practice, the melting index represents how much a given sequence differs from a CG-only sequence by taking into account the interaction potentials. The resulting melting indexes are then correlated to the experimental melting temperatures using linear regression. Further details concerning the model implementation are described in Refs. [45] and [44], and a free software implementation is reported in Ref. [46].

2.3. Optimisation of the model parameters

To optimise the PB model parameters we let them evolve from a predetermined starting point until we arrive at values that are able to minimise the square differences between experimental melting temperature T_i, and predicted melting temperature T_i' of a particular sequence i, for all N sequences of the dataset. This is represented as minimizing χ^2

$$\chi^2 = \sum_{i=1}^{N}(T_i' - T_i)^2. \tag{4}$$

We will also refer to the average melting temperature deviation ΔT

$$\langle \Delta T \rangle = \frac{1}{N}\sum_{i=1}^{N}|T_i' - T_i|. \tag{5}$$

The predicted melting temperature T_i' of a given sequence can be calculated from its melting index τ_i which is given by the PB model [45].

Unlike our previous calculations with DNA [22] and RNA [23], we could not implement a length dependence regression due to insufficient variability of sequence lengths (mostly 12 pb) in the dataset or Ref. [11]. Therefore, we employed a single linear regression equation, without length dependence,

$$T_i' = f_0 + f_1\tau_i, \tag{6}$$

where f_0 and f_1 are linear regression coefficients calculated for all sequences that belong to a particular mismatch, for instance I.A.

2.3.1. Minimisation procedure

The single-equation regression scheme, Eq. (6), generally presents poorer results than the two-equation regression [23] since

the best global minima of χ^2 is much more difficult to find. Therefore the minimisation procedure was adapted in the following way:

Initial minimisation All stacking and Morse potentials are allowed to vary simultaneously. For each minimisation round the initial parameters change randomly within the interval of 16.5 meV and 49.5 meV. This interval corresponds to a variation of ±50% of the Morse parameter of the AT base pair (D_{AT} = 33 meV). Similarly, all initial stacking parameters are values chosen in the range between 1.25 eV/nm^2 and 3.75 eV/nm^2. We repeated this procedure 1000 times, which takes around 600 h of processing time on 1.6 GHz processors. The resulting value of each parameter is obtained by averaging over all results provided by the minimisation.

Final minimisation We use the averaged results from the previous round as new initial parameters. But this time, instead of varying the initial parameters, the experimental dataset is randomly changed by small amounts according to the declared experimental error in the temperature measurements, 0.3 °C in this case [11]. In this way we can estimate the experimental error influence on our parameters [22]. As in the initial minimisation, all parameters are allowed to vary during minimisation with the goal of obtaining the smallest χ^2. This procedure was repeated again 1000 times. At the end we obtained both the standard deviation and the final optimum value for each parameter.

2.4. Availability

The inosine parameters calculated in this work were included in version 2.0 of our TfReg software [46], and can be found at https://sites.google.com/site/geraldweberufmg/tfreg

3. Results and discussion

3.1. Parameter optimisation

The present analysis differs from our previous calculations for DNA and RNA [22,23] in two essential ways. One is that it is an application of the Peyrard–Bishop model to non-canonical nucleotides and base-pairs that are mismatched, which to our knowledge has not been attempted before. The second is that the available inosine dataset contains no diversity of sequence lengths, which makes it impossible to apply the length dependent melting temperature regression developed for DNA [45]. Instead we have to recur to a single equation for the regression, Eq. (6), which was first introduced for an initial analysis of RNA [23]. This has resulted in a difficult convergence of χ^2 values during initial step of the minimisation procedure, which is the reason why a rather large number of minimisation steps was necessary. Fortunately, these limitations did not prevent us from obtaining fairly consistent results. After performing the two minimisation rounds, described in (2), we obtained an average temperature deviation $\langle \Delta T \rangle$ of 1.1 °C using the melting index technique [45]. Note that this is practically identical to the temperature deviation of 1.2 °C obtained from nearest-neighbour calculations reported by Watkins et al. [11].

Table 1
Inosine Morse potential parameters. The Morse potentials D are given in meV and λ in 10^{-2} nm. In parenthesis are the standard deviations, obtained from the final minimisation round, in compact uncertainty notation.

	IA	IC	IG	IT	II
D	31 (4)	34 (2)	38 (2)	29 (5)	24 (5)
λ	2.9 (9)	3.2 (2)	3.0 (3)	4.9 (2)	3.8 (1)

While the quality of prediction of the PB model and the nearest-neighbour calculations are comparable, the PB model offers a more intuitive interpretation of the model parameters. This is nicely illustrated in Table 1 where we present the resulting Morse potential parameters which are related to hydrogen bond strengths. The four inosine mismatches which pair with canonical nucleotides (A, C, G and T) resulted in Morse parameters similar to the Watson-Crick base pair AT of 33 meV [22]. This is consistent with the established view that two hydrogen bonds are formed for these inosine mismatches [11,2,15]. The potential widths λ for IA, IC and IG are mostly within the range of AT (3.2×10^{-2} nm). However, for IT and II we observe an important increase, Table 1, indicating a small widening of the potential.

The similarity of all four hydrogen bonds of inosine-canonical mismatches would support the suggested role of inosine as a universal base. However, the calculated stacking parameters clearly dispel this notion. A large part of these parameters fall outside the DNA range of 1 eV/nm^2 to 4 eV/nm^2 [22], as shown in Table 2. Only IC is reasonably contained within the canonical DNA range. At the extreme ends of the stacking parameter range is IG with the largest difference, 0.013–12.5 eV/nm^2, and IA with the second largest difference 0.0002–7.0 eV/nm^2. One explanation for the extreme stacking parameters of IG could be that it is the only mismatch

Table 2
Stacking parameters k in eV/nm^2. In parenthesis are the standard deviations in compact uncertainty notation. The table is divided into three parts, with the central section representing stacking parameters in the same range as for Watson–Crick base pairs [22].

Range	IA		IC		I.G		IT		I.I	
	ATpAI	0.0002[a]			ATpGI	0.013 (5)	ATpIT	0.24 (8)	IIpAT	0.09 (3)
	AIpAT	0.20 (6)			IGpAT	0.043 (9)	ATpTI	0.6 (2)	ATpII	0.8 (2)
	IApAT	0.9 (2)			GIpAT	0.12 (2)	ITpAT	0.9 (2)		
<1					CGpIG	0.69 (7)				
≥1			CGpCI	1.30 (9)			TApIT	1.5 (3)		
			CIpAT	1.8 (1)						
			CIpCG	1.57 (1)						
≥2	AIpTA	2.2 (3)	ATpCI	2.2 (1)	CGpGI	2.0 (2)	GCpIT	2.7 (4)	CGpII	2.6 (4)
	AIpGC	2.2 (3)	ATpIC	2.6 (2)			CGpTI	2.8 (4)		
			ICpAT	2.9 (2)						
≥3	CGpAI	3.4 (3)	GCpCI	3.4 (2)						
<4			CGpIC	3.8 (2)						
≥4	AIpCG	5.1 (4)			GCpIG	4.6 (3)	CGpIT	5.9 (4)	GCpII	4.6 (6)
	GCpAI	7.0 (5)			ATpIG	5.8 (4)	GCpTI	6.9 (4)		
					GCpGI	12.5 (4)				

[a] The standard deviation for this parameter is 0.004 eV/nm^2.

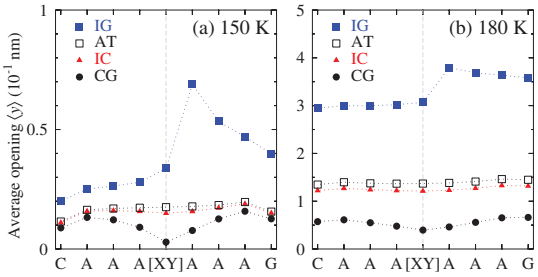


Fig. 1. Average opening profiles ⟨y⟩ as a function of base pair position of sequences of type CAAA[XY]AAAG, where XY is either IG (blue boxes), AT (squares), IC (red triangles) or CG (bullets). The position of the XY base pair is shown as a light grey vertical dashed line. Profiles were calculated at (a) 150 K and (b) 180 K. Please note that these temperatures are unrelated to the predicted melting temperature as discussed in Ref. [22]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this letter.).

in syn-anti configuration [11]. Rutledge et al. [19], who performed a density functional theory (DFT) analysis on hypoxanthine also reported considerable buckle and propeller distortions for guanine mismatches. On the other hand, IC is reasonably contained within Watson–Crick base pair range possibly because inosine is closest to guanine. In other words, IC stacking resembles more closely a Watson–Crick GC base pair than any other inosine mismatch, which is consistent with the similarity between inosine and guanine.

Nearest-neighbour (NN) parameters of free-energy (FE) models also present very large variations as the PB stacking parameters, in certain cases up to an order of magnitude [11]. For instance, for IA mismatches the FE parameter ΔG°_{37} are 0.09 kcal/mol and −1.24 kcal/mol, for AIpAT and AIpCG nearest-neighbours, respectively. Another example is for the case of II mismatches ΔG°_{37} of −0.09 kcal/mol and 0.81 kcal/mol for GCpII and IIpAT, respectively. However, correlating linear FE parameters to PB parameters is usually not meaningful [47] since PB parameters are non-linear and the model considers NN and base-pair interactions separately. Instead, we will discuss the overall stability of the inosine mismatches in terms of the average openings ⟨y⟩, Eq. (3), calculated with the new parameters. The average openings provide a view on sequence stability which differs from Gibbs free energy estimates. The value of ΔG indicates the likelihood that a given structure may hybridise spontaneously at a given temperature. In contrast, average openings ⟨y⟩ provide a more detailed view of helix denaturation by calculating the distance of base pairs along the sequence.

Figure 1 shows the calculated opening profiles ⟨y⟩ as a function of temperature for some inosine mismatches embedded in a CAAAIAAAG sequence. This sequence was used by Martin et al. [2] which is one of the earliest studies on inosine. A smaller average opening ⟨y⟩ at a given base pair means that the sequence is locally more stable at a given temperature. IC mismatches present an almost identical opening profile to AT base pairs, while IG in contrast destabilises the duplex quite strongly. Furthermore, the IG mismatch causes a much increased base pair amplitude on the right-hand side of the duplex, where there is a much weaker stacking interaction. This asymmetric disturbance propagates along all sites towards the right-hand side terminal, and remains even at higher temperatures.

IA mismatches also have similar variable stacking interactions as IG, shown in Table 2. In Figure 2 we show an example where an IA mismatch is placed at the 5th position of a 14 bp long sequence. Even at low temperatures the IA mismatch strongly disturbs the duplex with very large but localised average openings. Interestingly, at a somewhat larger temperature the average helix opening is much larger on the left-hand side despite being composed of strongly bonded CG base pairs. Adding further CG base pairs at the left-hand side dampens these amplitudes as shown in Figure 3a. Yet as we increase the temperature to 240 K, Figure 3b, even these

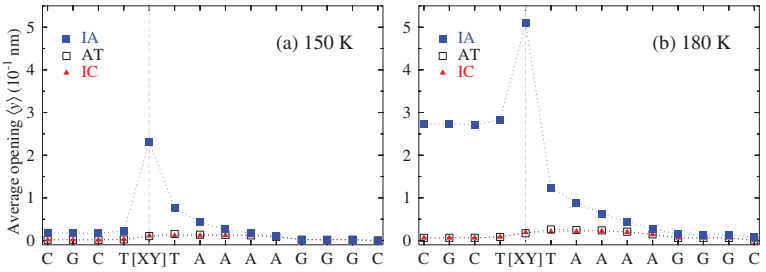


Fig. 2. Average opening profiles ⟨y⟩ as function of base pair position of sequences of type CGCT[XY]TAAAAGGGC, where XY is either IA (blue boxes), AT (squares) or IC (red triangles). The position of the XY base pair is shown as a light grey vertical dashed line. Profiles were calculated at (a) 150 K and (b) 180 K. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this letter.)

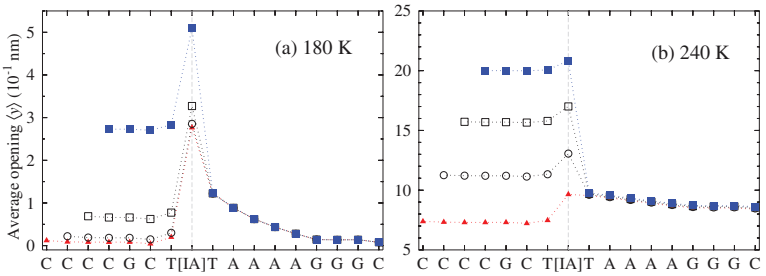


Fig. 3. Average opening profiles ⟨y⟩ as function of base pair position of sequences of increasing length with IA mismatch (light grey vertical dashed line) calculated at (a) 180 K and (b) 240 K. The first sequence is the same as of Figure 2 and shown as blue boxes. The three remaining sequences are as the first with one (squares), two (circles) or three (red triangles) additional CG base pairs at the left-hand side. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this letter.)

additional CG base pairs have little effect in dampening the motion amplitude. An intriguing aspect that emerges from Figure 3 is that the strong discontinuity of the stacking parameters essentially isolates each side of the sequences. The right-hand side is completely undisturbed by the additional CG base pairs on the left-hand side. Therefore, what we are observing appears to be a breakdown of the base pair cooperativity within the DNA sequence induced by the IA mismatch [48].

The cooperativity breakdown shown in Figures 2 and 3 suggest uses of technological interest for inosine mismatches beyond that of probe design. For instance, the difference in motion amplitude on each side of a sequence containing IG or IA mismatches should affect applications which depend on FRET, which is highly sensitive to donor–acceptor distances [49]. But perhaps more interestingly could be its use as thermal rectifier, a role that was suggested for normal DNA sequences [34,50].

4. Conclusions

In this work we have obtained the Peyrard–Bishop parameters, both for Morse potential and elastic constants for sequences containing inosine mismatches in DNA. Such parameters were calculated from published melting thermodynamics datasets and were used to predict melting temperatures and general stability trends for inosine mismatches. The calculated parameters which represent hydrogen bonding are in agreement with the notion of all inosine mismatches being formed by two hydrogen bonds. The stacking parameters are distributed over a much wider range than for canonical DNA base pairs. In some cases, such as IA and IG mismatches, these stacking parameters take either very low or very high values which causes a disruption in the melting cooperativity along the chain of DNA base pairs.

Acknowledgements

The authors acknowledge the agencies that funded this research: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Fapemig). RVM thanks Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for his doctoral fellowship.

Appendix A. Supplementary data

Supplementary data associated with this letter can be found, in the online version, at http://dx.doi.org/10.1016/j.cplett.2015.04.045

References

[1] Y. Takahashi, K. Kato, Y. Hayashizaki, T. Wakabayashi, E. Ohtsuka, S. Matsuki, M. Ikehara, K. Matsubara, Proc. Natl. Acad. Sci. 82 (7) (1985) 1931.
[2] F.H. Martin, M.M. Castro, F. Aboul-ela, I. Tinoco, Nucl. Acids Res. 13 (24) (1985) 8927.
[3] E. Ohtsuka, S. Matsuki, M. Ikehara, Y. Takahashi, K. Matsubara, J. Biol. Chem. 260 (5) (1985) 2605, http://arxiv.org/abs/http://www.jbc.org/cgi/reprint/260/5/2605.pdf, http://www.jbc.org/cgi/content/abstract/260/5/2605.
[4] G.M. Rossolini, S. Cresti, A. Ingianni, P. Cattani, M.L. Riccio, G. Satta, Mol. Cell. Probes 8 (2) (1994) 91.
[5] D. Loakes, Nucl. Acids. Res. 29 (12) (2001) 2437.
[6] A. Ho, M. Murphy, S. Wilson, S.R. Atlas, J.S. Edwards, BMC Genomics 12 (1) (2011) 598.
[7] Z. Yu, Engineering 5 (2013) 440.
[8] T. Baumann, K.M. Arndt, K.M. Müller, BMC Biotechnol. 13 (1) (2013) 81.
[9] Y. Kawase, S. Iwai, H. Inoue, K. Miura, E. Ohtsuka, Nucl. Acids Res. 14 (19) (1986) 7727.
[10] S.C. Case-Green, E.M. Southern, Nucl. Acids. Res. 22 (2) (1993) 131.
[11] J. Watkins, E. Norman, J. SantaLucia, John, Nucl. Acids. Res. 33 (19) (2005) 6258, http://dx.doi.org/10.1093/nar/gki918, http://arxiv.org/abs/http://nar.oxfordjournals.org/cgi/reprint/33/19/6258.pdf, http://nar.oxfordjournals.org/cgi/content/abstract/33/19/6258
[12] C. Öhrmalm, M. Jobs, R. Eriksson, S. Golbob, A. Elfaitouri, F. Benachenhou, M. Strømme, J. Blomberg, Nucl. Acids Res. 38 (21) (2010) e195, http://dx.doi.org/10.1093/nar/gkq777
[13] S. Uesugi, Y. Oda, M. Ikehara, Y. Kawase, E. Ohtsuka, J. Phys. C 262 (15) (1987) 6965, http://arxiv.org/abs/http://www.jbc.org/cgi/reprint/262/15/6965.pdf, http://www.jbc.org/cgi/content/abstract/262/15/6965.
[14] Y. Oda, S. Uesugi, M. Ikehara, Y. Kawase, E. Ohtsuka, Nucl. Acids Res. 19 (19) (1991) 5263.
[15] E.M. Basílio Janke, F. Riechert-Krause, K. Weisz, J. Phys. Chem. B 115 (26) (2011) 8569.
[16] P.W.R. Corfield, W.N. Hunter, T. Brown, P. Robinson, O. Kennard, Nucl. Acids Res. 15 (19) (1987) 7935.
[17] W.B.T. Cruse, J. Aymani, O. Kennard, T. Brown, A.G.C. Jack, G.A. Leonard, Nucl. Acids Res. 17 (1) (1989) 55.
[18] G. Paragi, I. Pálinkó, C. Van Alsenoy, I.K. Gyémánt, B. Penke, Z. Timár, New J. Chem. 26 (10) (2002) 1503.
[19] L.R. Rutledge, C.A. Wheaton, S.D. Wetmore, Phys. Chem. Chem. Phys. 9 (4) (2007) 497.
[20] L.R. Rutledge, S.D. Wetmore, Phys. Chem. Chem. Phys. 14 (8) (2012) 2743.
[21] M. Krepl, M. Otyepka, P. Banáš, J. Šponer, J. Phys. Chem. B 117 (6) (2013) 1872.
[22] G. Weber, J.W. Essex, C. Neylon, Nat. Phys. 5 (2009) 769, http://dx.doi.org/10.1038/nphys1371
[23] G. Weber, Nucl. Acids. Res. 41 (2013) e30, http://dx.doi.org/10.1093/nar/gks964 http://nar.oxfordjournals.org/content/early/2012/10/18/nar.gks964.short
[24] M. Peyrard, A.R. Bishop, Phys. Rev. Lett. 62 (23) (1989) 2755.
[25] J. Šponer, C.A. Morgado, D. Svozil, J. Phys. Chem. B 116 (28) (2012) 8331.
[26] C.A. Johnson, R.J. Bloomingdale, V.E. Ponnusamy, C.A. Tillinghast, B.M. Znosko, M. Lewis, J. Phys. Chem. B 115 (29) (2011) 9244.
[27] T.D. Amarante, G. Weber, J. Phys.: Conf. Ser. 490 (1) (2014) 012203.
[28] S. Buyukdagli, M. Joyeux, Chem. Phys. Lett. 484 (4) (2010) 315.
[29] A. Shigaev, O. Ponomarev, V. Lakhno, Chem. Phys. Lett. 513 (4) (2011) 276.
[30] G. Kalosakas, Phys. Rev. E 84 (2011) 051905, http://dx.doi.org/10.1103/PhysRevE.84.051905
[31] S. Behnia, A. Akhshani, M. Panahi, A. Mobaraki, M. Ghaderian, Phys. Lett. A 376 (2012) 2538.
[32] A.E.B. Pupo, F. Falo, A. Fiasconaro, J. Chem. Phys. 139 (9) (2013) 095101.
[33] A. Bergues-Pupo, J. Bergues, F. Falo, Phys. A: Stat. Mech. Appl. 396 (2014) 99.
[34] C.-C. Chien, K.A. Velizhanin, Y. Dubi, M. Zwolak, Nanotechnology 24 (9) (2013) 095704.
[35] H.-H. Huang, P. Lindblad, J. Biol. Eng 7 (10) (2013).
[36] L. Pettinato, E. Calistri, F. Di Patti, R. Livi, S. Luccioli, PloS One 9 (1) (2014) e85260.
[37] G. Weber, Europhys. Lett. 73 (5) (2006) 806, http://dx.doi.org/10.1209/epl/i2005-10466-6
[38] D.X. Macedo, I. Guedes, E.L. Albuquerque, Phys. A: Stat. Mech. Appl. 404 (2014) 234.
[39] A. Singh, N. Singh, Phys. A: Stat. Mech. Appl. 392 (9) (2013) 2052.
[40] R.F. Álvarez-Estrada, G.F. Calvo, H. Serrano, J. Comput. Appl. Math. 236 (15) (2012) 3561.
[41] C.B. Passos, P.S. Kuhn, M.C. Barbosa, Phys. A: Stat. Mech. Appl. 413 (2014) 481.
[42] D.E. Bergstrom, P. Zhang, W.T. Johnson, Nucl. Acids Res. 25 (10) (1997) 1935.
[43] Y.-L. Zhang, W.-M. Zheng, J.-X. Liu, Y.Z. Chen, Phys. Rev. E 56 (6) (1997) 7100.
[44] G. Weber, N. Haslam, J.W. Essex, C. Neylon, J. Phys. Condens. Matter 21 (2009), http://dx.doi.org/10.1088/0953-8984/21/3/034106
[45] G. Weber, N. Haslam, N. Whiteford, A. Prügel-Bennett, J.W. Essex, C. Neylon, Nat. Phys. 2 (2006) 55, http://dx.doi.org/10.1038/nphys189
[46] G. Weber, Bioinformatics 29 (2013) 1345, http://dx.doi.org/10.1093/bioinformatics/btt133 http://bioinformatics.oxfordjournals.org/content/29/10/1345
[47] G. Weber, Europhys. Lett. 96 (2011) 68001 http://iopscience.iop.org/0295-5075/96/6/68001
[48] S. Ares, N.K. Voulgarakis, K.Ø. Rasmussen, A.R. Bishop, Phys. Rev. Lett. 94 (2005) 035504.
[49] S. Khrapunov, M. Brenowitz, Biophys. J. 86 (1) (2004) 371.
[50] M. Terraneo, M. Peyrard, G. Casati, Phys. Rev. Lett. 88 (9) (2002) 094302.

Apêndice E

Artigo publicado no estudo de caso do câncer de tireóide MCT-MEN2-RET

PLOS ONE

RESEARCH ARTICLE

Effect of 3′UTR *RET* Variants on *RET* mRNA Secondary Structure and Disease Presentation in Medullary Thyroid Carcinoma

Lucieli Ceolin[1], Mirian Romitti[1], Débora Rodrigues Siqueira[1], Carla Vaz Ferreira[1], Jessica Oliboni Scapineli[1], Beatriz Assis-Brazil[2], Rodolfo Vieira Maximiano[3], Tauanne Dias Amarante[3], Miriam Celi de Souza Nunes[3], Gerald Weber[3], Ana Luiza Maia[1]*

1 Thyroid Section, Endocrine Division, Hospital de Clínicas de Porto Alegre, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil, **2** Pathology Department, Hospital de Clínicas de Porto Alegre, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil, **3** Department of Physics, Computational Biophysics Group, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

* almaia@ufrgs.br

CrossMark
click for updates

OPEN ACCESS

Citation: Ceolin L, Romitti M, Rodrigues Siqueira D, Vaz Ferreira C, Oliboni Scapineli J, Assis-Brazil B, et al. (2016) Effect of 3′UTR *RET* Variants on *RET* mRNA Secondary Structure and Disease Presentation in Medullary Thyroid Carcinoma. PLoS ONE 11(2): e0147840. doi:10.1371/journal.pone.0147840

Editor: H. Sunny Sun, National Cheng Kung University Medical College, TAIWAN

Received: June 12, 2015

Accepted: January 8, 2016

Published: February 1, 2016

Copyright: © 2016 Ceolin et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper.

Funding: This work was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Apoio a Pesquisa do Rio Grande do Sul (FAPERGS), and Fundo de Incentivo a Pesquisa do Hospital de Clínicas de Porto Alegre (FIPE), Brazil.

Abstract

Background

The *RET* S836S variant has been associated with early onset and increased risk for metastatic disease in medullary thyroid carcinoma (MTC). However, the mechanism by which this variant modulates MTC pathogenesis is still open to discuss. Of interest, strong linkage disequilibrium (LD) between *RET* S836S and 3'UTR variants has been reported in Hirschsprung's disease patients.

Objective

To evaluate the frequency of the *RET* 3'UTR variants (rs76759170 and rs3026785) in MTC patients and to determine whether these variants are in LD with S836S polymorphism.

Methods

Our sample comprised 152 patients with sporadic MTC. The *RET* S836S and 3'UTR (rs76759170 and rs3026785) variants were genotyped using Custom TaqMan Genotyping Assays. Haplotypes were inferred using the phase 2.1 program. *RET* mRNA structure was assessed by Vienna Package.

Results

The mean age of MTC diagnosis was 48.5±15.5 years and 57.9% were women. The minor allele frequencies of *RET* polymorphisms were as follows: S836S, 5.6%; rs76759170, 5.6%; rs3026785, 6.2%. We observed a strong LD among S836S and 3'UTR variants ($|D'| = -1$, $r^2 = 1$ and $|D'| = -1$, $r^2 = 0{,}967$). Patients harboring the S836S/3'UTR variants presented a higher percentage of lymph node and distant metastasis ($P = 0.013$ and $P<0.001$, respectively). Accordingly, RNA folding analyses demonstrated different RNA secondary

Competing Interests: The authors have declared that no competing interests exist.

structure predictions for WT(TCCGT), S836S(TTCGT) or 3'UTR(GTCAC) haplotypes. The S836S/3'UTR haplotype presented a greater number of double helices sections and lower levels of minimal free energy when compared to the wild-type haplotype, suggesting that these variants provides the most thermodynamically stable mRNA structure, which may have functional consequences on the rate of mRNA degradation.

Conclusion

The *RET* S836S polymorphism is in LD with 3'UTR variants. *In silico* analysis indicate that the 3'UTR variants may affect the secondary structure of *RET mRNA*, suggesting that these variants might play a role in posttranscriptional control of the *RET* transcripts.

Introduction

Medullary thyroid carcinoma (MTC), a malignant tumor originating in parafollicular C cells of the thyroid, represents about 4% of all thyroid cancers [1]. MTC is mainly sporadic (75%), but may also be part of an inherited disorder transmitted as an autosomal dominant trait with 100% penetrance, referred to as multiple endocrine neoplasia type 2 (MEN 2). The MEN 2 syndrome is classified into three distinct clinical subtypes: MEN type 2A (MEN 2A), characterized by the presence of MTC, pheochromocytoma (PHEO), and hyperparathyroidism (HPT); MEN type 2B (MEN 2B), that includes MTC, PHEO, ganglioneuromatosis, and marfanoid habitus; familial MTC (FMTC), characterized by MTC as the only feature of the disease [2,3].

The *RET* (REarranged during Transfection) proto-oncogene is the susceptibility gene for hereditary MTC and *RET* molecular analysis is now considered critical in MTC management [4,5]. Genotype–phenotype correlation has been reported and mutations in exons 5, 8, 10, 11, 14, 15 or 16 occur in more than 98% of MEN2 families [4,5]. The RET gene, a member of the cadherin superfamily, encodes a receptor tyrosine kinases, which are cell-surface molecules that transduce signals for cell growth and differentiation. This gene plays a crucial role in neural crest development, and it can undergo oncogenic activation *in vivo* and *in vitro* by cytogenetic rearrangement. Alternative splicing of the *RET* gene results in the production of *RET51*, *RET43* and *RET9* isoforms, contain 51, 43 and 9 amino acids in their C-terminal tail, respectively. The biological roles of isoforms RET51 and RET9 are the most well studied *in-vivo* as these are the most abundant isoforms in which *RET* occurs [6]. RET9 and RET51 isoforms appear to be coexpressed in MTC [7,8]. The RET 51 isoform with MEN2-related mutations might activate the Ras signaling cascade with a greater efficiency than the RET 9 isoform. Nevertheless, it has not been observed significant differences in the transforming ability of RET51 and RET9 [6].

Although much is known about hereditary MTC, the molecular mechanisms underlying sporadic MTC tumors remain to be clarified. Somatic RET point mutations in exon 16 (M918T) and gene deletions are described in 23–66% of sporadic MTC [9–11]. However, these mutations are not uniform among the various cell subpopulations, suggesting that the tumor may be of polyclonal origin or that the RET mutations are not the initial events in MTC tumorigenesis. Thus, in the last years, several studies have also investigated the relationship between the presence of single nucleotide polymorphisms (SNPs) in the *RET* gene and susceptibility or progression of sporadic MTC. The *RET* neutral variants L769L, S836S, and S904S have been associated with clinical presentation and disease outcome in sporadic MTC patients [12]. Particularly, the S836S variant (codon 836 of exon 14, SerAGC/SerAGT) has been found over-represented in sporadic MTC patients from Germany, Spain, and the United States [13,14]. This

variant also has been associated with two-to-three-fold increase in the risk of MTC in the Spanish population [14], early onset of MTC and increased risk for metastatic disease [15]. These results are in agreement with a recent meta-analysis that demonstrated an association between sporadic MTC susceptibility and S836S polymorphism [16]. Nevertheless, other studies failed to demonstrate any effect of *RET* variants on risk of development or on the natural course of MTC [17–19].

On the other hand, the mechanism by which SNPs modulate the MTC pathogenesis is still open to discussion [12,16]. One of the proposed mechanisms suggests that the S383S neutral variant might be in linkage disequilibrium (LD) with an unknown functional variant, which may modulate the *RET* oncogene expression by affecting the mRNA stability [13,20,21]. Recently, the presence of multiple *RET* risk alleles has been associated with increased estimated risk for MTC development and aggressiveness, indicating that these variants could be acting in an additive manner on disease pathogenesis [22].

The *RET* mutations are also associated with Hirschsprung's disease (HSCR), a developmental disorder of the enteric nervous system, characterized by the absence of ganglion cells in the distal colon resulting in functional obstruction [23]. However, contrasting with the gain-of function effect caused by *RET* mutations in MTC, loss-of-function *RET* mutations account for approximately 50% of familial and 7–35% of sporadic HSCR patients [24,25]. A potential explanation for these distinct effects is that constitutive RET activation is sufficient to induce neoplastic transformation of the C-cells and adrenal chromaffin cells, but not to produce a trophic response in the precursor neurons [26]. Of interest, *RET* polymorphisms and haplotypes have been described as underrepresented in HSCR patients, which could act as low susceptibility loci and modify the phenotype of HSCR [27–29]. Particularly, the haplotype including the uncommon S836S polymorphic allele has shown a low penetrant protective effect against the disease [30,31]. Interestingly, Griseri et al 2007 report a strong LD among S836S and 3'untranslated region (3'UTR) variants (rs76759170, G>A and rs3026785, T>C) [32]. The 3'UTR gene region is emerging as fundamental and effective in regulating gene expression at posttranscriptional levels (pre-mRNA processing, mRNA stability and translational regulation) [33], suggesting that 3'UTR variants might have a functional effect on the RET expression [32].

Here, we evaluated the presence of linkage disequilibrium between S836S and 3'UTR *RET* variants in MTC patients and assessed whether the 3'UTR variants could play a role **in** *RET* expression.

Material and Methods

Ethics Statement

The information obtained during the study did not affect the patients' diagnosis or treatment. The protocol was approved by the Committee on Research Ethics from Hospital de Clínicas de Porto Alegre (project number 12–0225), and all patients and/or their legal representatives provided written informed consent for the study. Clinical investigation was conducted according to the principles expressed in the Declaration of Helsinki.

Patients

Patients with a diagnosis of MTC attending the Endocrine Division at Hospital de Clínicas de Porto Alegre were invited to participate in the study. Since 1997, our division has been a reference center for *RET* mutation screening in Brazil, and therefore, patients referred to us by other Brazilian centers were also invited to participate.

The patients underwent a complete clinical examination, laboratory tests (levels of basal calcitonin (Until December 2003, Calcitonin IRMA-DSL7700, Diagnostic Systems Laboratories,

Inc., Webster, TX, USA, reference range <10 pg/ml and, after January 2004, Immulite 2000, Diagnostic Products Corporation, Los Angeles, CA, USA; reference value (VR) male <12.0 pg/ml and female <6.0 pg/ml)), plasma parathyroid hormone (PTH; Immulite 2000 Intact PTH, Diagnostic Products), urinary fractionated metanephrines (HPLC), and, whenever indicated, diagnostic imaging investigation (cervical ultrasonography, thorax and abdominal computed tomography (CT)). Selected patients were submitted to whole-body metaiodobenzylguanidine scintigraphy to rule out distant metastasis.

In our division, the MTC treatment follows the protocols recommended by the current guidelines [4,5]. Total thyroidectomy was performed in all patients with varying cervical neck dissection procedures. The diagnosis of lymph node metastasis was based on histological examination. Patients with suspicious distant metastasis (i.e. the presence of local metastasis and/or serum calcitonin >150 pg/ml) underwent imaging exams (cervical, thoracic and abdomen CT (or liver magnetic resonance imaging), and bone scintigraphy). Patients with undetectable calcitonin levels were considered free of disease. Tumor staging was performed according to the current International Union against Cancer TNM classification[34].

Genotyping Assay

RET variants L769L (rs1800861, codon 769, exon 13, LeuCTT/LeuCTG), S836S (rs1800862; codon 836, exon 14, SerAGC/SerAGT), S904S (rs1800863, codon 904, exon 15, SerTCC/SerTCG) and 3'UTR (rs3026785, T>C and rs76759170, G>A) were analyzed in DNA extracted from peripheral blood leukocytes by standardized procedures. Genotype analysis was performed using Human Custom TaqMan SNP Genotyping Assays (Applied Biosystems, Foster City, CA, USA), as described previously by [22,35]. Primer and probe sequences used for genotyping the *RET* 3'UTR variants were, rs3026785: 5′ -CACGTAAATGCAGAAGTTACTAAGTATTAAGTATTACT-3′ (forward primer), 5′ -AGGAACATGATCTGGTTTAATGACCTTT-3′ (reverse primer), VIC-5′ -TCTGTCAGTTATTAAAATT-3′ , FAM-5′ -TGTCAGTTACTAAAATT-3′ ; rs75759170: 5′ -ACACGTAACCTGGCTCTAATTTGG-3′ (forward primer), 5′ -CTGCATTTAGTAAGACTATCATTAAGCATATCTGA-3′ (reverse primer), VIC-5′ -CACAGTGTATCTGAAAAA-3′ , FAM-5′ -CACAGTGTATTTGAAAAA-3′ .

The reactions were conducted in a 96-well plate, in a total 5 ml reaction volume using 2-10ng genomic DNA, TaqMan Genotyping Master Mix 1X (Applied Biosystems), and Custom TaqMan Genotyping Assay 1X. The plates were then positioned in a real-time PCR thermal cycler (7500 Fast Real PCR System; Applied Biosystems) and heated for 10 min at 95°C, followed by 45 cycles of 95°C for 15s and 62°C for 1min. Fluorescence data files from each plate were analyzed using automated allele calling software (SDS 2.1; Applied Biosystems).

Haplotype Construction and Linkage Disequilibrium Analysis

The haplotypes were constructed based on the combination of allelic variants and their frequencies were inferred using the phase 2.1 program, which implements a Bayesian statistical method [36]. We also used the phase 2.1 program to compare the distributions of different *RET* haplotypes between MTC patients through permutation analyses of 1000 random replicates [36]. Among all pairs of biallelic loci, we examined widely used measures of linkage disequilibrium (LD), Lewontin's D' |D'| and r^2 [37].

Somatic M918T *RET* Mutation Analysis

Forty seven paraffin-embedded MTC samples were available by sequencing. Samples were sequenced at the Unidade de Análises Moleculares e de Proteínas (Centro de Pesquisa Experimental, HCPA) using ABI 3500 Genetic Analyzer with 50 cm capillaries and POP7 polymer

(Applied Biosystems). PCR products were labeled with 5.0 pmol of the primer 5′ -AGGGATAG GGCCTGGGCTTC-3′ and 1 μL of BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) in a final volume of 10 μL. Labeling reactions were performed in a Veriti® 96-Well Thermal Cycler (Applied Biosystems) with an initial denaturing step of 96°C for 1 min followed by 35 cycles of 96°C for 15 sec, 50°C for 15 sec and 60°C for 4 min. Labeled samples were purified using BigDye XTerminator Purification Kit (Applied Biosystems) and electroinjected in the automatic sequencer.

Thermodynamic Simulations Methodology

The bioinformatics analyses were realized based on haplotypes inferred by Phase program. Additionally, we also used the haplotype carrying only the S836S variant (TTCGT). No patient harbored only S836S polymorphism was observed in our study population, this variant occurs only in combination with L769L and 3'UTR variants (GTCAC haplotype).

The preliminary analysis was realized in both, wild type (TCCGT_WT haplotype) and polymorphic sequences (GCCGT, TCGGT and GTCAC haplotypes), with the RNAfold algorithm provided by the Vienna Package 2.1.5. This algorithm provides the optimal RNA secondary structure and its associated minimal free energy (MFE) for each input sequence [38].

The Vienna Package [38] program RNAsubopt with default parameters was used to generate suboptimal secondary structures of all haplotypes. Since the number of possible suboptimal structures for each sequence is exceedingly large we randomly selected a subset of those using the -p option. In this analysis we used subsets of 2900 structures for each sequence. The resulting dot-bracket files were evaluated as follows: 1) by counting the number of continuous base-pairs in double helices of at least 6 bp (number of double helices sections—N_{DH}); 2) by calculating the resulting Gibbs free energy (kcal/mol) with RNAeval from the Vienna Package [38].

First we performed these calculations for the wild type sequence and obtained average numbers of base-pairs, total melting index and Gibbs free energies. Then we compared each suboptimal structure of the polymorphic haplotype to those averages and counted how many times they present higher thermodynamic stability (lower Gibbis free energy). This gives us a measure that represents the probability which polymorphic haplotype may result into a more stable structure than the wild-type sequence.

The NCBI Reference Sequence accession number for the *RET* gene was NM_020975.4 (mRNA).

Immunohistochemistry Analysis

Immunohistochemistry analysis (IHC) was performed on thin sections (3 μm) of previously formalin-fixed and paraffin-embedded tissues. The antibody used was monoclonal rabbit anti-human RET (ab134100; Abcam Inc., Cambridge, MA, USA), Sections representing MTC were submitted to routine immunohistochemical technique, which comprises deparaffination and rehydration, antigenic recovery, inactivation of endogenous peroxidase, and blockage of unspecific reactions. Primary antibodies were incubated overnight at a temperature of 4°C, at dilutions of 1:400 followed by application of streptavidin horseradish peroxidase conjugate (LSAB; DakoCytomation, Via Real Carpinteria, CA, USA), and diaminobenzidine tetrahydrochloride (Kit DAB; DakoCytomation). Absence of the primary antibody was used as a negative control.

Semi-Quantitative Analysis for the Intensity of Positive Staining in Tissues

The intensity of positive staining of RET was performed by digital image analysis using the Image-Pro Plus 6.0 software (Media Cybernetics, Rockville, MD, USA). Two independent

researchers (L.C. and J.O.S) carried out blind analysis of the intensity of brownish-colored immunostaining in pixels in 4 to 12 fields from each slide, according of the tumor size area. The mean number of pixels identified by both researchers in each sample was used. All of the images were taken using the same microscope and camera sets [39].

Statistical Analysis

Results are expressed as mean ± S.D. or median (IQ 25–75) unless otherwise specified. Hardy–Weinberg equilibrium for each polymorphism was assessed by χ^2 tests. Baseline characteristics were compared using χ^2 tests or Fisher's exact test for qualitative variables. Quantitative variables were compared between groups using Student's t-test or Mann-Whitney test. The differences in cumulative lymph node and/or distant metastasis between groups were tested by Kaplan–Meier curves; comparisons between curves were performed using the log rank test. The Statistical Package for the Social Sciences 18.0 (SPSS Inc., Chicago, IL, USA) was used, and $P<0.05$ was considered as statistically significant.

Results

RET Variants in MTC Patients

The clinical and oncological features of the study subjects are summarized in Table 1. The median age at diagnosis was 48.5±15.5 years and the percentage of women was 57.9%. The ethnic background was 95% Caucasian. The median basal serum calcitonin and CEA levels at diagnosis were 782 (167–2566) pg/ml and 33.6 (7.13–115.4) ng/ml, respectively. The median tumor size was 2.3 (1.5–3.35) cm. Seventy-eight patients (58.3%) presented lymph node and 23.1% presented distant metastasis at diagnosis, respectively. The minor allele frequencies observed were as follows: S836S (5.6%) and 3'UTR *RET* variants rs76759170 (5.6%) and rs3026785 (6.2%) (Table 2).

Haplotype Construction and Linkage Disequilibrium Analysis

We used a Bayesian statistical method to estimate the frequency of different haplotypes produced by the combination of the L769L(rs1800861), S836S(rs1800862), S904S(rs1800863) and 3'UTR(rs3026785 and rs76759170) polymorphisms in MTC patients.

A total of 5 haplotypes were inferred in our study population. The haplotype frequencies were shown in Table 3. Interestingly, the S836S polymorphism occurs only in combination with L769L and 3'UTR variants. We observed a strong LD among S836S and 3'UTR variants:

Table 1. The clinical and oncological features of medullary thyroid carcinoma patients.

Total patients (%)	152
Sex female (%)	88 (57.9)
Age at Diagnosis (yr) [1]	48.5±15.5
Calcitonin (pg/ml) [2]	782 (167–2566)
CEA (ng/ml) [2]	33.6 (7.13–115.4)
Tumor Size (cm) [2]	2.3 (1.5–3.35)
N1 (%)	77 (58.3)
M1 (%)	27 (23.1)

CEA, carcinoembryonic antigen; N1, lymph node metastasis; M1, distant metastasis.

[1]Data expressed as mean±S.D.

[2]Data expressed as median (IQ 25–75).

doi:10.1371/journal.pone.0147840.t001

Table 2. *RET* minor allele frequency polymorphisms in medullary thyroid carcinoma patients.

RET Polymorphisms	Allele variation	Sporadic MTC (n = 152)
S836S (%)	C>T	5.6
rs76759170 (%)	G>A	5.6
rs3026785 (%)	T>C	6.2

doi:10.1371/journal.pone.0147840.t002

S836S and rs76759170 (|D'| = -1, r^2 = 1), S836S and rs3026785 (|D'| = -1, r^2 = 0.967), and rs76759170 and rs3026785 (|D'| = -1, r^2 = 0.967).

RET Haplotypes and Disease Presentation

Seventeen patients (11.2%) were heterozygous for S836S/3'UTR polymorphic allele (GTCAC haplotype). Individuals harboring the S836S/3'UTR haplotype (GTCAC) had higher serum calcitonin level (1860(1160–7087) *vs* 579(120–2014) pg/mL, P = 0.014), CEA level (584(145–4353) *vs* 26(7–113) ng/mL, P = 0.007) and larger tumor size (3.3(1.8–4.0) *vs* 2.1(1.4–3.0), P = 0.062). The S836S/3'UTR *RET* haplotype (GTCAC) was associated with higher rates of metastatic disease as compared to patients without GTCAC haplotype (87.5 *vs* 53.9%, P = 0.013 and 60 *vs* 16.8%, P = 0.001) (Table 4). Accordingly, Kaplan–Meier estimates of cumulative lymph node and distant metastasis yielded distinct curves for patients with or without the S836S/3'UTR haplotype (P = 0.011 and P<0.001, respectively; Fig 1), further demonstrating that metastatic disease occurred earlier in those individuals harboring the S836S/3'UTR variants.

Somatic M918T *RET* Mutation

The somatic *RET* M918T mutation has been previously associated with lymph node metastases at diagnosis and could be a confounding factor in our analysis. Thus we also evaluated somatic *RET* mutations in our sample population. Forty-seven paraffin-embedded samples were available. Of them, we failed **to** extract / amplify DNA from 7 even after several repeated attempts. Of the 40 DNA samples available for analysis, 13 (32.5%) samples have somatic M918T mutation. There were no significant difference in the frequency of somatic M918T mutation in samples with or without S836S/3ÚTR variants (10/31 (32.3%) vs. 3/9 (33.3%), P = 1.00; Fisher's exact test). In addition, no association was observed between the presence of somatic M918T mutation with lymph node or distant metastasis (P = 0.614 and P = 0.628, respectively).

Table 3. Haplotypes inferred by the Phase Program.

	Presence/absence of					HaplotypeFrequency[1]
Haplotypes	L769L	S836S	S904S	rs76759170	rs3026785	
TCCGT (WT)						55
GCCGT	X					18
TCGGT			X			21
GCCGC*	X				X	0.3
GTCAC	X	X		X	X	5.6

Haplotype frequencies was calculated by the Phase 2.0 program using permutation test (1000 replications). For simplified representation, the Phase program shows the haplotypes by the presence of the wild-type or risk alleles.

[1] Frequencies calculated by the Phase 2.0 program based on the number of chromosomes.

*The GCCGC haplotype was present in only two patients, thus was not included in our analyzes.

doi:10.1371/journal.pone.0147840.t003

Table 4. Clinical and oncological features of sporadic medullary thyroid carcinoma patients according to the presence of polymorphic haplotype.

	Without GTCAC	GTCAC	*P*
Total patients (%)*	133 (88.6)	17 (11.3)	
Sex female (%)	79 (59.4)	8 (47.1)	0.332
Age at Diagnosis (yr)	48.6±16	45.6±13	0.441[a]
Calcitonin (pg/ml)[1]	579 (120–2014)	1860 (1160–7087)	0.014[b,d]
CEA (ng/ml)[2]	26 (7–113)	584 (145–4353)	0.007[b,d]
Size Tumor (cm)[3]	2.1 (1.4–3.0)	3.3 (1.8–4)	0.062[b]
N1 (%)[4]	62 (53.9)	14 (87.5)	0.013[c]
M1 (%)[5]	17 (16.8)	9 (60)	0.001[c]

PHEO, pheochromocytoma; HPT, hyperparathyroidism; N1, lymph node metastasis; M1, distant metastasis; GTCAC, haplotype harboring L769L, S836S and 3'UTR *RET* variants; Without GTCAC, all other haplotype.

* The total of patients available was 150, the two patients carrying the GCCGC were not included in our analyzes.

[1] Data evaluated for 67 patients

[2] Data evaluated for 39 patients

[3] Data evaluated for 99 patients

[4] Data evaluated for 131 patients

[5] Data evaluated for 116 patients

[a]Data expressed as mean±S.D.

[b]Data expressed as median (IQ 25–75).

[c]Variables were compared using the Yates' X^2-test or Fisher's exact test.

[d]Variables were compared using the Mann–Whitney U test.

doi:10.1371/journal.pone.0147840.t004

In Silico Analysis of *RET* Haplotypes

Our purpose was to evaluate the effect of *RET* neutral variants on the full *RET* mRNA structure (5.6kb). In order to minimize the experimental limitations, increase the accuracy of computer predictions of RNA secondary structure and perform a more robust analysis, we evaluated the effect of the *RET* haplotypes on the optimal and suboptimal mRNA structures. The bioinformatics analyses used all haplotypes inferred by Phase program (Table 3) and included only synonymous polymorphisms.

As shown in Fig 2, RNA folding analyses have demonstrated different RNA secondary structure predictions for WT (TCCGT), S836S (TTCGT) and S836S/3'UTR (GTCAC) haplotypes. We can observe that S836S/3'UTR (GTCAC) haplotype presented a structural modification not observed in WT (TCCGT) or S836S (TTCGT) haplotypes.

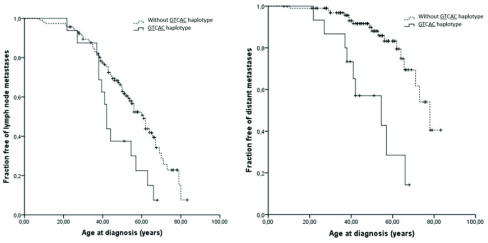


Fig 1. Kaplan–Meier estimates the proportion of sporadic MTC patients with lymph node (A; n = 131; P = 0.011) or distant metastasis (B; n = 116; P<0.001). The log rank test was used to compare curves.

doi:10.1371/journal.pone.0147840.g001

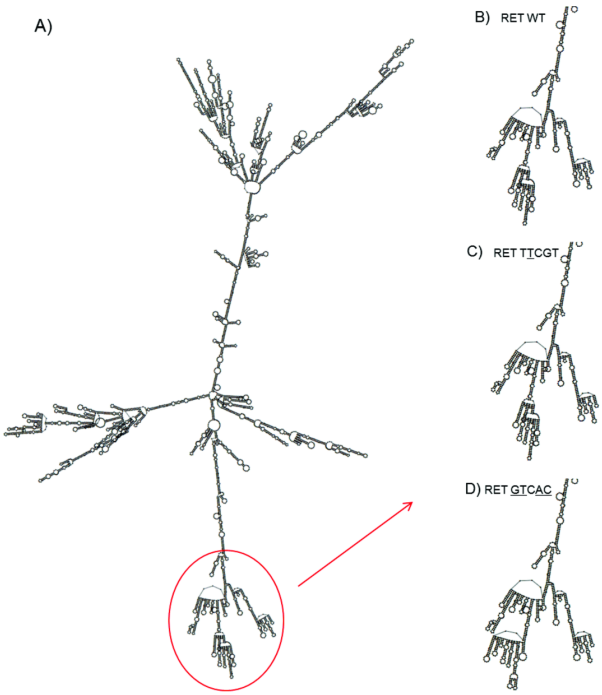


Fig 2. The figure shows the optimal mRNA secondary structure of the RET haplotypes. A) Total mRNA secondary structure of the *RET* wild-type (WT) haplotype. B) Part of mRNA secondary structure of the *RET* WT haplotype. C) Part of mRNA secondary structure of the *RET* S836S haplotype (TTCGT). D) Part of mRNA secondary structure of the *RET* 3'UTR haplotype (GTCAC). Haplotypes generated by RNAfold program (Vienna Package).

doi:10.1371/journal.pone.0147840.g002

Concerning the N_{DH} analyses, the S836S/3'UTR (GTCAC) haplotype presented greater N_{DH} in both, optimal and suboptimal structures when compared with the WT haplotype (65 *vs* 64, Fig 3B and 58 *vs* 57, Fig 3D, respectively). In agreement, the sequence carrying S836S/3'UTR (GTCAC) haplotype presented lower levels of MFE when compared to the WT sequence (optimal, -2142.1 *vs* -2138.6, Fig 3A; suboptimal, -1989.4 *vs* -1983.1, Fig 3C). Thus, the structure carrying the 3'UTR variants provides the most thermodynamically stable mRNA secondary structure when compared to the other.

Immunohistochemistry Analysis

Next, we decided to look whether the presence of the S836S/3'UTR (GTCAC) haplotype was associated with increased expression of RET protein in paraffin-embedded samples of patients with or without this haplotype. To minimize confounding factors, we selected only samples negative for somatic M918T mutation, since studies have demonstrated association between the presence of this mutation and stronger RET immunostaining [40].

Four samples carrying S836S/3'UTR (GTCAC) and 13 samples without the variants were available for analysis. Immunostaining for RET were detected in 100% of the samples analyzed. The overall median of the RET immunoreactivity observed was 220.2 (155–393) pixels. Of interest, all MTC samples carrying S836S/3'UTR (GTCAC) haplotype uniformly display a very strong intensity of RET immunostaining (Fig 4A). Remarkably, despite the relative small

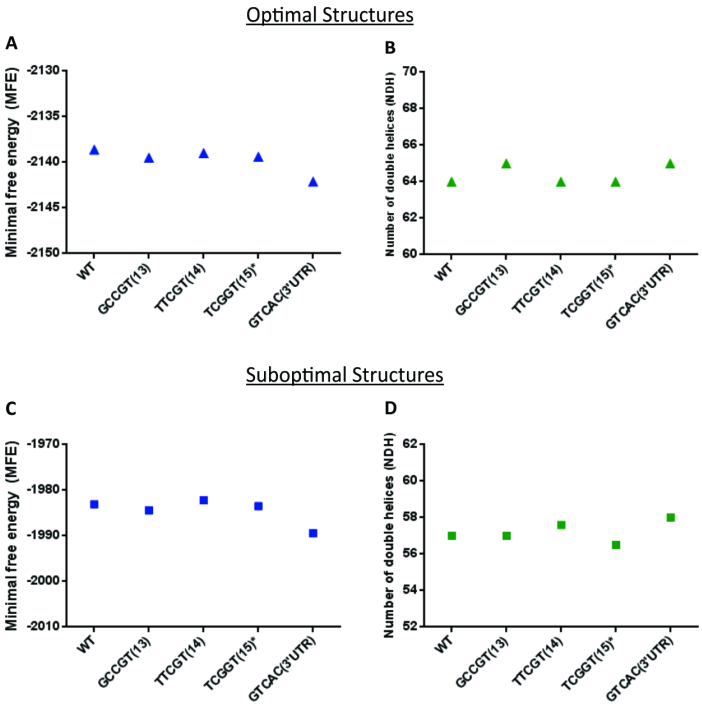


Fig 3. The minimal free energy (MFE, kcal/mol) and number of double helices (N_{DH}) were available in both, optimal and suboptimal structures. For suboptimal structures MFE and NDH are averages over 2900 samples. The variant fragment carrying the S836S and 3'UTR variants (GTCAC haplotype) presented greater N_{DH} (B, D) and lower levels of MFE (A, C) when compared to wild-type haplotype (WT, TCCGT), this fact happens in both, optimal and suboptimal structures. *These analysis included only synonymous polymorphisms.

doi:10.1371/journal.pone.0147840.g003

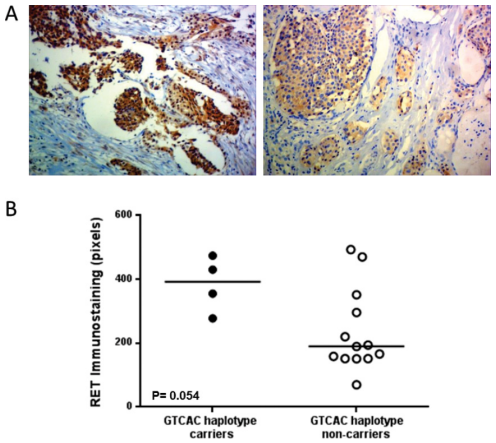


Fig 4. Immunostaining of the RET proto-oncogene in GTCAC haplotype carriers & non-carriers. A) Two representative slices of RET Immunostaining in a sample carrier S836S/3'UTR (GTCAC haplotype) (left) and non-carrier of this haplotype (right). B) Intensity of RET Immunostaining in samples with or without S836S/3'UTR (GTCAC haplotype), P = 0.054.

doi:10.1371/journal.pone.0147840.g004

number of samples, we observed a strong trend towards a significant statistical difference in the RET immunostaining between samples carrying S836S/3'UTR (GTCAC) haplotype and those without these variants (392.7 (297–463) vs. 189.5 (152–324), P = 0.054; Fig 4B).

Discussion

In the present study, we have demonstrated that the neutral *RET* S836S polymorphism is in linkage disequilibrium with the *RET* 3'UTR variants (rs3026785 and rs76759170) in patients with sporadic MTC. Interestingly, the individuals harboring these variants presented more aggressive disease and metastasis at younger age. Furthermore, the haplotype carrying the S836S/3'UTR *RET* variants presented a greater N_{DH} and lower levels of MFE as compared to the wild-type haplotype, which may have functional consequences on the rate of mRNA degradation. Interestingly, MTC samples carrying S836S/3'UTR (GTCAC) haplotype displayed stronger RET immunostaining than those without these variants, although the difference did not reach statistical significance.

Based on the overrepresentation of *RET* polymorphisms in individuals with MTC, a role for these genetic variants in the MTC pathogenesis has been suggested. However, the molecular mechanism by which *RET* S836S polymorphism affects the development or clinical course of MTC is still not properly understood. It has been suggested that polymorphisms could influence the *RET* mRNA levels. However, quantitative studies in MTC tumor samples show no difference in *RET* mRNA levels in patients with or without the S836S polymorphisms [41]. Another hypothesis suggests that bases exchange in the DNA molecule could create a new alternative splicing site, leading to the synthesis of a truncated protein, erroneous ligand binding, microRNA (miRNA) binding, change of structure and mRNA stability as well as a number of copies [42]. However, the S836S polymorphic variant does not affect DNA–protein binding or RNA splicing and editing of the *RET* gene [43]. An alternative way to explain the role of *RET* S836S polymorphisms on MTC proposes that this neutral variant can be in LD with an unknown functional variant [44].

Of interest, a strong LD between S836S polymorphism and 3'UTR variants has been reported in HSCR disease [32]. The 3'UTR gene region is emerging as fundamental and effective in regulating gene expression at posttranscriptional levels (pre-mRNA processing, mRNA stability and translational regulation) [33]. Several sequence elements that may be involved in mRNA regulation exist in the 3′UTR region, including regions rich in adenine and uridine elements—AU-rich elements (ARE). The ARE mRNAs are regulated by RNA-binding proteins that can selectively bind to the ARE and promote their mRNA stability or degradation [45]. Elegant studies performed by Griseri et al [31,32] demonstrated that the rs3026785 is located next to the AUUUA sequence and showed that a single nucleotide substitution may influence the secondary structure of *RET* mRNA decreasing *RET* mRNA degradation in human neuroblastoma cells, leading to increases of transcription product and, probably, in the amount of total *RET* protein at the cell membrane [31,32].

The S836S polymorphism is associated with increased RET expression in both diseases. Thus, interestingly, in HSCR, the S836S variant is associated with a protective effect since loss-of-function RET is causative of the disease [31,32]. We have previously shown that the *RET* S836S polymorphism is overrepresented in sporadic MTC patients [22]. Here, we show that the *RET* S836S polymorphism is in linkage disequilibrium with *RET* 3'UTR variants in MTC patients. Of interest, patients harboring these variants present higher levels of serum calcitonin and CEA, larger tumor size, and higher rates of metastatic disease as compared to those harboring other haplotypes. The somatic *RET* M918T mutation has been associated with advanced disease at diagnosis [9,46] and could be a confounding factor in our analysis. However, we

found no significant difference in the frequency of somatic M918T mutation in samples from patients with or without S836S/3UTR variants.

To further explore the modulating effect of *RET* S836S/3'UTR polymorphisms, we thought to evaluate the effects of the 3'UTR variants on the mRNA structure. Thermodynamics-based computational methods for prediction of RNA secondary structure are widely used to advance our understanding of the regulatory roles of cellular RNA [47,48]. These methods use MFE principles to find the RNA secondary structure from the set of all possible structures for a given RNA sequence. In this study, *in silico* analysis showed that the S836S/3'UTR haplotype (GTCAC) could influence the stability of the *RET* mRNA. We observed that this haplotype presents the highest N_{DH} and lower levels of MFE, suggesting that *RET* mRNA carrying the S836S/3'UTR haplotype provides the most thermodynamically stable mRNA structure, when compared to the others. Thus, our data suggest that the S836S/3'UTR haplotype may affect the secondary structure of *RET* mRNA, supporting the hypothesis of a functional involvement of the 3'UTR variant allele in *RET* mRNA stability. Based on the results presented here, we speculated that S836S variant promotes an increased activity of RET by increasing the mRNA survival, contributing for gain-of-function RET that is causative of MTC. Interestingly, immunohistochemical analysis shows that samples carrying S836S/3'UTR (GTCAC) haplotype displayed stronger RET staining intensity than those without these variants (Fig 4). Although the difference did not reach statistical significance, these observations may indicate an effect of the polymorphic haplotype on RET expression.

Potential miRNAs recognition site are identified in 3' UTR regions, thus, it is reasonable to speculate that the 3'UTR polymorphism could be involved in a ligation site for a given miRNA, interfering the *RET* expression. Therefore, we used the TargetScan and miRDB softwares to look for potential miRNA targets in the 3'UTR of *RET* gene. We can observe that the *RET* 3'UTR variants (rs3026758 and rs76759170) are situated near to several miRNA-binding sites, but none of them is situated in the miRNA seed region. The miRNA-binding sites for the hsa-miR-27a, hsa-miR-27b, hsa-miR-128a, hsa-miR-216a-3p and hsa-miR-3681-3p are situated near to rs76759170, although this variant is not in the miRNA seed region. Similar results were found to the hsa-miR-590-3p and the rs3026785. Additional studies are warranted to further explore this issue.

Conclusion

Our results demonstrate linkage disequilibrium between *RET* S836S and 3'UTR genetic variants. The patients harboring S836S/3'UTR haplotype presents more aggressive disease as compared to those harboring the other haplotypes. Furthermore, the *RET* mRNA sequence carrying the S836S/3'UTR haplotype present higher structural and thermodynamic stability, supporting the hypothesis of a functional involvement of the 3'UTR variant allele in the post-transcriptional control of a *RET* transcripts

Author Contributions

Conceived and designed the experiments: ALM GW LC. Performed the experiments: LC CVF JOS RVM TDA MCSN BAB. Analyzed the data: LC MR DRS BAB. Contributed reagents/materials/analysis tools: ALM GW. Wrote the paper: ALM GW LC.

References

1. Davies L, Welch HG (2006) Increasing incidence of thyroid cancer in the United States, 1973–2002. JAMA 295: 2164–2167. PMID: 16684987

2. Clark JR, Fridman TR, Odell MJ, Brierley J, Walfish PG, Freeman JL (2005) Prognostic variables and calcitonin in medullary thyroid cancer. Laryngoscope 115: 1445–1450. PMID: 16094121
3. Pelizzo MR, Boschin IM, Bernante P, Toniato A, Piotto A, Pagetta C, et al. (2007) Natural history, diagnosis, treatment and outcome of medullary thyroid cancer: 37 years experience on 157 patients. Eur J Surg Oncol 33: 493–497. PMID: 17125960
4. Maia AL, Siqueira DR, Kulcsar MA, Tincani AJ, Mazeto GM, Maciel LM (2014) Diagnosis, treatment, and follow-up of medullary thyroid carcinoma: recommendations by the Thyroid Department of the Brazilian Society of Endocrinology and Metabolism. Arq Bras Endocrinol Metabol 58: 667–700. PMID: 25372577
5. Wells SA Jr., Asa SL, Dralle H, Elisei R, Evans DB, Gagel RF, et al. (2015) Revised American Thyroid Association Guidelines for the Management of Medullary Thyroid Carcinoma The American Thyroid Association Guidelines Task Force on Medullary Thyroid Carcinoma. Thyroid.
6. Richardson DS, Rodrigues DM, Hyndman BD, Crupi MJ, Nicolescu AC, Mulligan LM (2012) Alternative splicing results in RET isoforms with distinct trafficking properties. Mol Biol Cell 23: 3838–3850. doi: 10.1091/mbc.E12-02-0114 PMID: 22875993
7. Myers SM, Eng C, Ponder BA, Mulligan LM (1995) Characterization of RET proto-oncogene 3' splicing variants and polyadenylation sites: a novel C-terminus for RET. Oncogene 11: 2039–2045. PMID: 7478523
8. Santoro M, Rosati R, Grieco M, Berlingieri MT, D'Amato GL, de Franciscis V, et al. (1990) The ret proto-oncogene is consistently expressed in human pheochromocytomas and thyroid medullary carcinomas. Oncogene 5: 1595–1598. PMID: 1701232
9. Elisei R, Cosci B, Romei C, Bottici V, Renzini G, Molinaro E, et al. (2008) Prognostic significance of somatic RET oncogene mutations in sporadic medullary thyroid cancer: a 10-year follow-up study. J Clin Endocrinol Metab 93: 682–687. PMID: 18073307
10. Eng C, Mulligan LM, Smith DP, Healey CS, Frilling A, Raue F, et al. (1995) Mutation of the RET protooncogene in sporadic medullary thyroid carcinoma. Genes Chromosomes Cancer 12: 209–212. PMID: 7536460
11. Marsh DJ, Learoyd DL, Andrew SD, Krishnan L, Pojer R, Richardson AL, et al. (1996) Somatic mutations in the RET proto-oncogene in sporadic medullary thyroid carcinoma. Clin Endocrinol (Oxf) 44: 249–257.
12. Ceolin L, Siqueira DR, Romitti M, Ferreira CV, Maia AL (2012) Molecular Basis of Medullary Thyroid Carcinoma: The Role of RET Polymorphisms. Int J Mol Sci 13: 221–239. doi: 10.3390/ijms13010221 PMID: 22312249
13. Gimm O, Neuberg DS, Marsh DJ, Dahia PL, Hoang-Vu C, Raue F, et al. (1999) Over-representation of a germline RET sequence variant in patients with sporadic medullary thyroid carcinoma and somatic RET codon 918 mutation. Oncogene 18: 1369–1373. PMID: 10022819
14. Ruiz A, Antinolo G, Fernandez RM, Eng C, Marcos I, Borrego S (2001) Germline sequence variant S836S in the RET proto-oncogene is associated with low level predisposition to sporadic medullary thyroid carcinoma in the Spanish population. Clin Endocrinol (Oxf) 55: 399–402.
15. Siqueira DR, Romitti M, da Rocha AP, Ceolin L, Meotti C, Estivalet A, et al. (2010) The RET polymorphic allele S836S is associated with early metastatic disease in patients with hereditary or sporadic medullary thyroid carcinoma. Endocr Relat Cancer 17: 953–963. doi: 10.1677/ERC-09-0312 PMID: 20801952
16. Figlioli G, Landi S, Romei C, Elisei R, Gemignani F (2012) Medullary thyroid carcinoma (MTC) and RET proto-oncogene: mutation spectrum in the familial cases and a meta-analysis of studies on the sporadic form. Mutat Res 752: 36–44. doi: 10.1016/j.mrrev.2012.09.002 PMID: 23059849
17. Berard I, Kraimps JL, Savagner F, Murat A, Renaudin K, Nicolli-Sire P, et al. (2004) Germline-sequence variants S836S and L769L in the RE arranged during Transfection (RET) proto-oncogene are not associated with predisposition to sporadic medullary carcinoma in the French population. Clin Genet 65: 150–152. PMID: 14984475
18. Sharma BP, Saranath D (2011) RET gene mutations and polymorphisms in medullary thyroid carcinomas in Indian patients. J Biosci 36: 603–611. PMID: 21857107
19. Wohllk N, Soto E, Bravo M, Becker P (2005) Polimorfismos G691S, L769L y S836S del proto-oncogen RET no se asocian a mayor riesgo de cáncer medular tiroideo esporádico en pacientes chilenos. Rev Méd Chile 133.
20. Fernandez RM, Pecina A, Antinolo G, Navarro E, Borrego S (2006) Analysis of RET polymorphisms and haplotypes in the context of sporadic medullary thyroid carcinoma. Thyroid 16: 411–417. PMID: 16646689

21. Tamanaha R, Camacho CP, Pereira AC, da Silva AM, Maciel RM, Cerutti JM (2009) Evaluation of RET polymorphisms in a six-generation family with G533C RET mutation: specific RET variants may modulate age at onset and clinical presentation. Clin Endocrinol (Oxf) 71: 56–64.
22. Ceolin L, Siqueira DR, Ferreira CV, Romitti M, Maia SC, Leiria L, et al. (2012) Additive effect of RET polymorphisms on sporadic medullary thyroid carcinoma susceptibility and tumor aggressiveness. Eur J Endocrinol 166: 847–854. doi: 10.1530/EJE-11-1060 PMID: 22345297
23. Heanue TA, Pachnis V (2007) Enteric nervous system development and Hirschsprung's disease: advances in genetic and stem cell studies. Nat Rev Neurosci 8: 466–479. PMID: 17514199
24. Attie T, Edery P, Lyonnet S, Nihoul-Fekete C, Munnich A (1994) [Identification of mutation of RET proto-oncogene in Hirschsprung disease]. C R Seances Soc Biol Fil 188: 499–504. PMID: 7780793
25. Lyonnet S, Edery P, Mulligan LM, Pelet A, Dow E, Abel L, et al. (1994) [Mutations of RET proto-oncogene in Hirschsprung disease]. C R Acad Sci III 317: 358–362. PMID: 8000915
26. Asai N, Jijiwa M, Enomoto A, Kawai K, Maeda K, Ichiahara M, et al. (2006) RET receptor signaling: dysfunction in thyroid cancer and Hirschsprung's disease. Pathol Int 56: 164–172. PMID: 16634961
27. Fitze G, Appelt H, Konig IR, Gorgens H, Stein U, Walther W, et al. (2003) Functional haplotypes of the RET proto-oncogene promoter are associated with Hirschsprung disease (HSCR). Hum Mol Genet 12: 3207–3214. PMID: 14600022
28. Griseri P, Bachetti T, Puppo F, Lantieri F, Ravazzolo R, Devoto M, et al. (2005) A common haplotype at the 5' end of the RET proto-oncogene, overrepresented in Hirschsprung patients, is associated with reduced gene expression. Hum Mutat 25: 189–195. PMID: 15643606
29. Lantieri F, Griseri P, Puppo F, Campus R, Martucciello G, Ravazzolo R, et al. (2006) Haplotypes of the human RET proto-oncogene associated with Hirschsprung disease in the Italian population derive from a single ancestral combination of alleles. Ann Hum Genet 70: 12–26. PMID: 16441254
30. Fitze G, Cramer J, Ziegler A, Schierz M, Schreiber M, Kuhlisch E, et al. (2002) Association between c135G/A genotype and RET proto-oncogene germline mutations and phenotype of Hirschsprung's disease. Lancet 359: 1200–1205. PMID: 11955539
31. Griseri P, Pesce B, Patrone G, Osinga J, Puppo F, Sancandi M, et al. (2002) A rare haplotype of the RET proto-oncogene is a risk-modifying allele in hirschsprung disease. Am J Hum Genet 71: 969–974. PMID: 12214285
32. Griseri P, Lantieri F, Puppo F, Bachetti T, Di Duca M, Ravazzolo R, et al. (2007) A common variant located in the 3'UTR of the RET gene is associated with protection from Hirschsprung disease. Hum Mutat 28: 168–176. PMID: 16986122
33. Moore MJ (2005) From birth to death: the complex lives of eukaryotic mRNAs. Science 309: 1514–1518. PMID: 16141059
34. O'Sullivan B, Shah J (2003) New TNM staging criteria for head and neck tumors. Semin Surg Oncol 21: 30–42. PMID: 12923914
35. Siqueira DR, Ceolin L, Ferreira CV, Romitti M, Maia SC, Maciel LM, et al. (2014) Role of RET genetic variants in MEN2-associated pheochromocytoma. Eur J Endocrinol 170: 821–828. doi: 10.1530/EJE-14-0084 PMID: 24616415
36. Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68: 978–989. PMID: 11254454
37. Hedrick PW (1987) Gametic disequilibrium in cats: epistasis is an unlikely cause. J Hered 78: 204. PMID: 3611717
38. Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. (2011) ViennaRNA Package 2.0. Algorithms Mol Biol 6: 26. doi: 10.1186/1748-7188-6-26 PMID: 22115189
39. Brondani LA, de Souza BM, Duarte GC, Kliemann LM, Esteves JF, Marcon AS, et al. (2012) The UCP1 -3826A/G polymorphism is associated with diabetic retinopathy and increased UCP1 and MnSOD2 gene expression in human retina. Invest Ophthalmol Vis Sci 53: 7449–7457. doi: 10.1167/iovs.12-10660 PMID: 23033381
40. Eng C, Thomas GA, Neuberg DS, Mulligan LM, Healey CS, Houghton C, et al. (1998) Mutation of the RET proto-oncogene is correlated with RET immunostaining in subpopulations of cells in sporadic medullary thyroid carcinoma. J Clin Endocrinol Metab 83: 4310–4313. PMID: 9851769
41. Elisei R, Cosci B, Romei C, Bottici V, Sculli M, Lari R, et al. (2004) RET exon 11 (G691S) polymorphism is significantly more frequent in sporadic medullary thyroid carcinoma than in the general population. J Clin Endocrinol Metab 89: 3579–3584. PMID: 15240649
42. Borrego S, Saez ME, Ruiz A, Gimm O, Lopez-Alonso M, Antinolo G, et al. (1999) Specific polymorphisms in the RET proto-oncogene are over-represented in patients with Hirschsprung disease and may represent loci modifying phenotypic expression. J Med Genet 36: 771–774. PMID: 10528857

43. Griseri P, Sancandi M, Patrone G, Bocciardi R, Hofstra R, Ravazzolo R, et al. (2000) A single-nucleotide polymorphic variant of the RET proto-oncogene is underrepresented in sporadic Hirschsprung disease. Eur J Hum Genet 8: 721–724. PMID: 10980580

44. Borrego S, Wright FA, Fernandez RM, Williams N, Lopez-Alonso M, Davuluri R, et al. (2003) A founding locus within the RET proto-oncogene may account for a large proportion of apparently sporadic Hirschsprung disease and a subset of cases of sporadic medullary thyroid carcinoma. Am J Hum Genet 72: 88–100. PMID: 12474140

45. Lopez de Silanes I, Quesada MP, Esteller M (2007) Aberrant regulation of messenger RNA 3'-untranslated region in human cancer. Cell Oncol 29: 1–17.

46. Romei C, Elisei R, Pinchera A, Ceccherini I, Molinaro E, Mancusi F, et al. (1996) Somatic mutations of the ret protooncogene in sporadic medullary thyroid carcinoma are not restricted to exon 16 and are associated with tumor recurrence. J Clin Endocrinol Metab 81: 1619–1622. PMID: 8636377

47. Ding Y (2006) Statistical and Bayesian approaches to RNA secondary structure prediction. RNA 12: 323–331. PMID: 16495231

48. Lou F, Clote P (2010) Thermodynamics of RNA structures by Wang-Landau sampling. Bioinformatics 26: i278–286. doi: 10.1093/bioinformatics/btq218 PMID: 20529917

Referências Bibliográficas

[1] Francis S Collins, Ari Patrinos, Elke Jordan, Aravinda Chakravarti, Raymond Gesteland, LeRoy Walters, et al. New goals for the us human genome project: 1998-2003. *science*, 282 (5389):682–689, 1998.

[2] Martin Egli e Wolfram Saenger. *Principles of nucleic acid structure*. Springer Science & Business Media, 2013.

[3] Le Cong, F Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D Hsu, Xuebing Wu, Wenyan Jiang, Luciano A Marraffini, et al. Multiplex genome engineering using CRISPR/Cas systems. *Science*, 339(6121):819–823, 2013.

[4] Thomas Gaj, Charles A Gersbach, e Carlos F Barbas III. Zfn, talen, and crispr/cas-based methods for genome engineering. *Trends in biotechnology*, 31(7):397–405, 2013.

[5] MT Record Jr, SJ Mazur, P Melancon, JH Roe, SL Shaner, e L Unger. Double helical dna: conformations, physical properties, and interactions with ligands. *Annual review of biochemistry*, 50(1):997–1024, 1981.

[6] Stephen Neidle. *Principles of nucleic acid structure*. Academic Press, 2010.

[7] File:base pair gc.svg - wikimedia commons. `https://commons.wikimedia.org/wiki/File:Base_pair_GC.svg`, . (Acessado em 10/26/2017).

[8] File:base pair at.svg - wikimedia commons. `https://commons.wikimedia.org/wiki/File:Base_pair_AT.svg`, . (Acessado em 10/26/2017).

[9] Yaru Wang, Na Ma, Yan Wang, e Guangju Chen. Allosteric analysis of glucocorticoid receptor-dna interface induced by cyclic py-im polyamide: a molecular dynamics simulation study. *PloS one*, 7(4):e35159, 2012.

[10] Carl Schildkraut e Shneior Lifson. Dependence of the melting temperature of DNA on salt concentration. *Biopoly.*, 3(2):195–208, 1965.

[11] G Felsenfeld, David R Davies, e Alexander Rich. Formation of a three-stranded polynucleotide molecule. *J. Am. Chem. Soc.*, 79(8):2023–2024, 1957.

[12] Gary Felsenfeld e Alexander Rich. Studies on the formation of two-and three-stranded polyribonucleotides. *Biochimica et Biophysica Acta*, 26(3):457–468, 1957.

[13] Alexander Rich. DNA comes in many forms. *Gene*, 135(1):99–109, 1993.

[14] Helena Guiset Miserachs, Daniela Donghi, Richard Börner, Silke Johannsen, e Roland KO Sigel. Distinct differences in metal ion specificity of RNA and DNA G-quadruplexes. *JBIC Journal of Biological Inorganic Chemistry*, 21(8):975–986, 2016.

[15] Xiang-Jun Lu e Wilma K Olson. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, 31(17): 5108–5121, 2003.

[16] File:base stack.png. `http://www.tesi.mintrada.com/node80.html`. (Acessado em 10/26/2017).

[17] chemistryfam: Oligo-nucleotide synthesis. `http://www.ramisnazir.com/2016/12/oligo-nucleotide-synthesis.html`. (Acessado em 11/02/2017).

[18] Robert T Batey, Robert P Rambo, Jennifer A Doudna, et al. Tertiary motifs in RNA structure and folding. *Angewandte Chemie International Edition*, 38(16):2326–2343, 1999.

[19] Wilma K. Olson, Manju Bansal, Stephen K. Burley, Richard E. Dickerson, Mark Gerstein, Stephen C. Harvey Udo Heinemann, Xiang-Jun Lu, Stephen Neidle, Zippora Shakked Heinz Sklenar, Masashi Suzuki, Chang-Shung Tung, Eric Westhof Cynthia Wolberger, e Helen M. Berman. A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.*, 313:229–237, 2001.

[20] Neocles B Leontis e Eric Westhof. Geometric nomenclature and classification of RNA base pairs. *RNA*, 7(4):499–512, 2001.

[21] ChulHee Kang, Xiaohua Zhang, Robert Ratliff, Robert Moyzis, e Alexander Rich. Crystal structure of four-stranded oxytricha telomeric dna. *Nature*, 356(6365):126–131, 1992.

[22] Nucleic acid structure. `http://www.atdbio.com/content/5/Nucleic-acid-structure`. (Accessed on 29/11/2017).

[23] Kathleen P Howard. Thermodynamics of DNA duplex formation: A biophysical chemistry laboratory experiment. *J. Chem. Educ*, 77(11):1469, 2000.

[24] Sherrie Schreiber-Gosche e Robert A Edwards. Thermodynamics of oligonucleotide duplex melting. *Journal of Chemical Education*, 86(5):644, 2009.

[25] John SantaLucia, Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA*, 95(4):1460–1465, 1998. URL `http://www.pnas.org/cgi/content/abstract/95/4/1460`.

[26] Gerald Weber. Optimization method for obtaining nearest-neighbour DNA entropies and enthalpies directly from melting temperatures. *Bioinformatics*, 31(6):871–877, 2015. doi: 10.1093/bioinformatics/btu751. URL `http://bioinformatics.oxfordjournals.org/content/31/6/871`.

[27] K. J. Breslauer, R Frank, H Blocker, e L. A. Marky. Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. USA*, 83(11):3746–3750, 1986.

[28] John SantaLucia, Jr., H T Allawi, e P A Seneviratne. Improved nearest-neighbour parameters for predicting DNA duplex stability. *Biochem.*, 35:3555–3562, 1996.

[29] Tianbing Xia, John SantaLucia, Jr., Mark E. Burkard, Ryszard Kierzek, Susan J. Schroeder, Xiaoqi Jiao, Christopher Cox, e Douglas H. Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochem.*, 37:14719–14735, 1998.

[30] M. Peyrard e A. R. Bishop. Statistical mechanics of a nonlinear model for DNA denaturation. *Phys. Rev. Lett.*, 62(23):2755–2757, 1989.

[31] Lesley R Rutledge, Craig A Wheaton, e Stacey D Wetmore. A computational characterization of the hydrogen-bonding and stacking interactions of hypoxanthine. *Physical Chemistry Chemical Physics*, 9(4):497–509, 2007.

[32] Miroslav Krepl, Michal Otyepka, Pavel Banáš, e Jiří Šponer. Effect of guanine to inosine substitution on stability of canonical DNA and RNA duplexes: molecular dynamics thermodynamics integration study. *The Journal of Physical Chemistry B*, 117(6):1872–1879, 2013.

[33] C Bouchiat, MD Wang, J-F Allemand, T Strick, SM Block, e V Croquette. Estimating the persistence length of a worm-like chain molecule from force-extension measurements. *Biophysical journal*, 76(1):409–413, 1999.

[34] Hans C Fogedby e Ralf Metzler. Dynamics of dna breathing: weak noise analysis, finite time singularity, and mapping onto the quantum coulomb problem. *Physical Review E*, 76 (6):061915, 2007.

[35] Michel Peyrard, Santiago Cuesta-López, e Guillaume James. Modelling DNA at the mesoscale: a challenge for nonlinear science? *Nonlinearity*, 21(6):T91, 2008.

[36] Husin Alatas e Dede Hermanudin. Semi-discrete DNA breather in Peyrard-Bishop-Dauxois model with fifth-order-approximation Morse potential. *Chaos, Solitons & Fractals*, 45(9): 1231–1238, 2012.

[37] Titus S. van Erp, Santiago Cuesta-Lopez, Johannes-Geert Hagmann, e Michel Peyrard. Can one predict DNA transcription start sites by studying bubbles? *Phys. Rev. Lett.*, 95(21): 218104, 2005. URL `http://link.aps.org/abstract/PRL/v95/e218104`.

[38] S Behnia, A Akhshani, M Panahi, A Mobaraki, e M Ghaderian. Multifractal analysis of thermal denaturation based on the Peyrard-Bishop-Dauxois model. *Phys. Rev. E*, 84(3):031918, 2011.

[39] Navin Singh e Yashwant Singh. Statistical theory of force-induced unzipping of DNA. *The European Physical Journal E*, 17(1):7–19, 2005.

[40] T. Das e S. Chakraborty. A generalized Langevin formalism of complete dna melting transition. *Europhys. Lett.*, 83:48003, 2008.

[41] Marco Zoli. Stacking interactions in denaturation of DNA fragments. *Eur. Phys. J. E*, 34(7): 1–10, 2011.

[42] Marco Zoli. Anharmonic stacking in supercoiled DNA. *J. Phys.: Condens. Matter*, 24(19): 195103, 2012.

[43] Julien Barré e Thierry Dauxois. Lyapunov exponents as a dynamical indicator of a phase transition. *EPL (Europhysics Letters)*, 55(2):164, 2001.

[44] R. F. Machado e G. Weber. Wavelet coefficients as a guide to DNA phase transitions. *Europhys. Lett.*, 87(3):38005, 2009. doi: 10.1209/0295-5075/96/68001.

[45] Maria Barbi, Simona Cocco, e Michel Peyrard. Helicoidal model for DNA opening. *Phys. Lett. A*, 253:358–369, 1999.

[46] Simona Cocco e Remi Monasson. Statistical mechanics of torque induced denaturation of DNA. *Phys. Rev. Lett.*, 83:5178–81, 1999.

[47] S. Cuesta-López, M. Peyrard, e D. J. Graham. Model for DNA hairpin denaturation. *Eur. Phys. J. E*, 16:235–246, 2005.

[48] Gerald Weber, Niall Haslam, Jonathan W. Essex, e Cameron Neylon. Thermal equivalence of DNA duplexes for probe design. *J. Phys.: Condens. Matter*, 21:034106, 2009. doi: 10.1088/0953-8984/21/3/034106.

[49] D. X. Macedo, I. Guedes, e E. L. Albuquerque. Thermal properties of a DNA denaturation with solvent interaction. *Physica A: Statistical Mechanics and its Applications*, 404:234–241, 2014. doi: 10.1016/j.physa.2014.02.029.

[50] Philip M Morse. Diatomic molecules according to the wave mechanics. II. Vibrational levels. *Physical Review*, 34(1):57, 1929.

[51] Ramón F Álvarez-Estrada, Gabriel F Calvo, e Hélia Serrano. A transfer integral technique for solving a class of linear integral equations: Convergence and applications to DNA. *Journal of Computational and Applied Mathematics*, 236(15):3561–3571, 2012.

[52] Yong-Li Zhang, Wei-Mou Zheng, Ji-Xing Liu, e Y. Z. Chen. Theory of DNA melting based on the Peyrard-Bishop model. *Phys. Rev. E*, 56(6):7100–7115, 1997.

[53] T. Dauxois, M. Peyrard, e A. R. Bishop. Entropy-driven DNA denaturation. *Phys. Rev. E*, 47 (1):R44–R47, 1993.

[54] T. Dauxois e M. Peyrard. Entropy-driven transition in a one-dimensional system. *Phys. Rev. E*, 51(5):4027–4040, 1995.

[55] Gerald Weber. Sharp DNA denaturation due to solvent interaction. *Europhys. Lett.*, 73(5): 806–811, 2006. doi: 10.1209/epl/i2005-10466-6.

[56] Nikos Theodorakopoulos, Thierry Dauxois, e Michel Peyrard. Order of the phase transition in models of DNA thermal denaturation. *Phys. Rev. Lett.*, 85(1):6–9, 2000.

[57] Gerald Weber, Niall Haslam, Nava Whiteford, Adam Prügel-Bennett, Jonathan W. Essex, e Cameron Neylon. Thermal equivalence of DNA duplexes without melting temperature calculation. *Nat. Phys.*, 2:55–59, 2006. doi: 10.1038/nphys189.

[58] Richard Owczarzy, Yong You, Bernardo G. Moreira, Jeffrey A. Manthey, Lingyan Huang, Mark A. Behlke, e Joseph A. Walder. Effects of sodium ions on DNA duplex oligomers: Improved predictions of melting temperatures. *Biochem.*, 43:3537–3554, 2004.

[59] Gerald Weber, Jonathan W. Essex, e Cameron Neylon. Probing the microscopic flexibility of DNA from melting temperatures. *Nat. Phys.*, 5:769–773, 2009. doi: 10.1038/nphys1371.

[60] Gerald Weber. Mesoscopic model parametrization of hydrogen bonds and stacking interactions of RNA from melting temperatures. *Nucleic Acids Res.*, 41:e30, 2013. doi: 10.1093/nar/gks964. URL `http://nar.oxfordjournals.org/content/41/1/e30`.

[61] M. Swart, C. F. Guerra, e F. M. Bickelhaupt. Hydrogen bonds of RNA are stronger than those of DNA, but NMR monitors only presence of methyl substituent in uracil/thymine. *J. Am. Chem. Soc.*, 126(51):16718–16719, 2004.

[62] Rodolfo Vieira Maximiano e Gerald Weber. Deoxyinosine mismatch parameters calculated with a mesoscopic model result in uniform hydrogen bonding and strongly variable stacking interactions. *Chem. Phys. Lett.*, 631–632:87–91, 2015. doi: 10.1016/j.cplett.2015.04.045.

[63] Ross B Inman. Transitions of DNA homopolymers. *J. Mol. Biol.*, 9(3):624–637, 1964.

[64] RB Inman e RL Baldwin. Helix-random coil transitions in DNA homopolymer pairs. *Journal of Molecular Biology*, 8(4):452–469, 1964.

[65] Yoosuke Takahashi, Kikuya Kato, Yoshihide Hayashizaki, Toshiaki Wakabayashi, Eiko Ohtsuka, Shigeru Matsuki, Morio Ikehara, e Kenichi Matsubara. Molecular cloning of the human cholecystokinin gene by use of a synthetic probe containing deoxyinosine. *Proceedings of the National Academy of Sciences*, 82(7):1931–1935, 1985.

[66] Francis H Martin, Miguel M Castro, Fareed Aboul-ela, e Ignacio Tinoco. Base pairing involving deoxyinosine: implications for probe design. *Nucleic Acids Research*, 13(24):8927–8938, 1985.

[67] Gian Maria Rossolini, Stefania Cresti, Angela Ingianni, Paola Cattani, Maria Letizia Riccio, e Giuseppe Satta. Use of deoxyinosine-containing primers vs degenerate primers for polymerase chain reaction based on ambiguous sequence information. *Molecular and Cellular Probes*, 8(2):91–98, 1994.

[68] D. Loakes. The application of universal DNA base analogues. *Nucleic Acids Res.*, 29(12): 2437–2447, 2001.

[69] Tobias Baumann, Katja M Arndt, e Kristian M Müller. Directional cloning of DNA fragments using deoxyinosine-containing oligonucleotides and endonuclease V. *BMC Biotechnology*, 13(1):81, 2013.

[70] Antoine Ho, Maurice Murphy, Susan Wilson, Susan R Atlas, e Jeremy S Edwards. Sequencing by ligation variation with endonuclease v digestion and deoxyinosine-containing query oligonucleotides. *BMC Genomics*, 12(1):598, 2011.

[71] Zhiliang Yu. Substitution of deoxyinosine as universal base in oligonucleotides for DNA ligation. *Engineering*, 5:440, 2013.

[72] Y Kawase, S Iwai, H Inoue, K Miura, e E Ohtsuka. Studies on nucleic acid interactions i. stabilities of mini-duplexes (dG2A4XA4G2.dC2T4YT4C2) and self-complementary d(GGGAAXYTTCCC) containing deoxyinosine and other mismatched bases. *Nucleic Acids Research*, 14(19):7727–7736, 1986.

[73] S. C. Case-Green e E. M. Southern. Studies of the base pairing properties of deoxyinosine by solid phase hybridisation to oligonucleotides. *Nucleic Acids Res.*, 22(2):131–136, 1993.

[74] Jr Watkins, Norman E. e Jr SantaLucia, John. Nearest-neighbor thermodynamics of deoxyinosine pairs in DNA duplexes. *Nucleic Acids Res.*, 33(19):6258–6267, 2005. doi: 10.1093/nar/gki918. URL `http://nar.oxfordjournals.org/cgi/content/abstract/33/19/6258`.

[75] Christina Öhrmalm, Magnus Jobs, Ronnie Eriksson, Sultan Golbob, Amal Elfaitouri, Farid Benachenhou, Maria Strømme, e Jonas Blomberg. Hybridization properties of long nucleic acid probes for detection of variable target sequences, and development of a hybridization prediction algorithm. *Nucleic Acids Research*, 38(21):e195, 2010. doi: 10.1093/nar/gkq777.

[76] S Uesugi, Y Oda, M Ikehara, Y Kawase, e E Ohtsuka. Identification of I:A mismatch base-pairing structure in DNA. *J. Phys. C:*, 262(15):6965–6968, 1987. URL `http://www.jbc.org/cgi/content/abstract/262/15/6965`.

[77] Thomas Hupp, Christian Sturm, Eline M Basílio Janke, Montserrat Pérez Cabre, Klaus Weisz, e Bernd Engels. A combined computational and experimental study of the hydrogen-bonded dimers of xanthine and hypoxanthine. *J. Phys. Chem. A*, 109(8):1703–1712, 2005.

[78] Peter W R Corfield, William N Hunter, Tom Brown, Paul Robinson, e Olga Kennard. Inosine.adenine base pairs in a B-DNA duplex. *Nucleic Acids Research*, 15(19):7935–7949, 1987.

[79] Gábor Paragi, István Pálinkó, Cris Van Alsenoy, Iván K Gyémánt, Botond Penke, e Zoltán Timár. Ab initio studies on the H-bonding of hypoxanthine and DNA bases. *New Journal of Chemistry*, 26(10):1503–1506, 2002.

[80] Lesley R Rutledge e Stacey D Wetmore. A computational proposal for the experimentally observed discriminatory behavior of hypoxanthine, a weak universal nucleobase. *Physical Chemistry Chemical Physics*, 14(8):2743–2753, 2012.

[81] Donald E Bergstrom, Peiming Zhang, e W Travis Johnson. Comparison of the base pairing properties of a series of nitroazole nucleobase analogs in the oligodeoxyribonucleotide sequence 5'-d(CGCXAATTYGCG)-3'. *Nucleic Acids Research*, 25(10):1935–1942, 1997.

[82] W. B. T. Cruse, J. Aymani, Olga Kennard, Tom Brown, Audrey G. C. Jack, e G. A. Leonard. Refined crystal structure of an octanucleotide duplex with IT mismatched base pairs. *Nucleic Acids Research*, 17(1):55–72, 1989.

[83] Yasushi Oda, Seiichi Uesugi, Morio Ikehara, Yasutoshi Kawase, e Eiko Ohtsuka. NMR studies for identification of dI:dG mismatch base-pairing structure in DNA. *Nucleic Acids Research*, 19(19):5263–5267, 1991.

[84] Valerie Faibis, Jean AH Cognet, Yves Boulard, Lawrence C Sowers, e G Victor Fazakerley. Solution structure of two mismatches GG and II in the K-ras gene context by nuclear magnetic resonance and molecular dynamics. *Biochemistry*, 35(46):14452–14464, 1996.

[85] A. Campa e A. Giansanti. Experimental tests of the Peyard-Bishop model applied to the melting of very short DNA chains. *Phys. Rev. E*, 58(3):3585–3588, 1998.

[86] Peptide syn - transfer rna - wikipedia. `https://en.wikipedia.org/wiki/Transfer_RNA/media/File:Peptide_syn.png`, . (Acessado em 17/10/2017).

[87] Homo sapiens mitochondrion, complete genome - nucleotide - ncbi. `https://www.ncbi.nlm.nih.gov/nuccore/NC_012920.1`. (Acessado em 20/10/2017).

[88] Susan M Huse, Les Dethlefsen, Julie A Huber, David Mark Welch, David A Relman, e Mitchell L Sogin. Exploring microbial diversity and taxonomy using ssu rrna hypervariable tag sequencing. *PLoS genetics*, 4(11):e1000255, 2008.

[89] Carl R Woese, Otto Kandler, e Mark L Wheelis. Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579, 1990.

[90] Department of biology | biology | memorial university of newfoundland. `https://www.mun.ca/biology/`. (Acessado em 10/20/2017).

[91] Escherichia coli o157:h7 strain sakai 23s ribosomal rna gene, complete - nucleotide - ncbi. `https://www.ncbi.nlm.nih.gov/nuccore/NR_076322.1`. (Acessado em 10/20/2017).

[92] Charles Lydeard, Wallace E Holznagel, Murray N Schnare, e Robin R Gutell. Phylogenetic analysis of molluscan mitochondrial lsu rdna sequences and secondary structures. *Molecular phylogenetics and evolution*, 15(1):83–102, 2000.

[93] Estrutura dos ribossomos. `http://labs.icb.ufmg.br/lbcd/grupo6/estrutura.html`. (Acessado em 10/20/2017).

[94] Harvey Lodish, Arnold Berk, S Lawrence Zipursky, Paul Matsudaira, David Baltimore, e James Darnell. Molecular cell biology 4th edition. *National Center for Biotechnology InformationÕs Bookshelf*, 2000.

[95] L Mori, MC Arias, CY Miyaki, e EMB Dessen. Código genético. *Revista da Sociedade Brasileira de Genética*, 15(1):25–, 2009.

[96] File:trna-phe yeast en.svg - wikimedia commons. `https://commons.wikimedia.org/wiki/File:TRNA-Phe_yeast_en.svg`, . (Acessado em 10/21/2017).

[97] David Rearick, Ashwin Prakash, Andrew McSweeny, Samuel S Shepard, Larisa Fedorova, e Alexei Fedorov. Critical association of ncrna with introns. *Nucleic acids research*, 39(6): 2357–2366, 2010.

[98] National human genome research institute (nhgri). `https://www.genome.gov/`. (Acessado em 10/23/2017).

[99] Lioudmila V Sharova, Alexei A Sharov, Timur Nedorezov, Yulan Piao, Nabeebi Shaik, e Minoru SH Ko. Database for mrna half-life of 19 977 genes obtained by dna microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA research*, 16(1): 45–58, 2008.

[100] Yimei Cai, Xiaomin Yu, Songnian Hu, e Jun Yu. A brief review on the mechanisms of mirna regulation. *Genomics, proteomics & bioinformatics*, 7(4):147–154, 2009.

[101] D.P. Bartel. MicroRNAs genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, 2004.

[102] Microrna research takes aim at cholesterol | nih director's blog. `https://directorsblog.nih.gov/2013/11/26/microrna-research-takes-aim-at-cholesterol/`. (Acessado em 10/23/2017).

[103] S Wolfsheimer e A K Hartmann. Minimum-free-energy distribution of RNA secondary structures: Entropic and thermodynamic properties of rare events. *Physical Review E*, 82(2): 021902, 2010.

[104] Richa Tiwari, Lucy Haque, Sutanwi Bhuiya, e Suman Das. Third strand stabilization of poly (U).poly(A)*poly(U) triplex by the naturally occurring flavone luteolin: A multi-spectroscopic approach. *International Journal of Biological Macromolecules*, 103:692–700, 2017.

[105] Shawn M Lyons, Dorota Gudanis, Stephen M Coyne, Zofia Gdaniec, e Pavel Ivanov. Identification of functional tetramolecular RNA G-quadruplexes derived from transfer RNAs. *Nature Communications*, 8(1):1127, 2017.

[106] Ruth Nussinov, George Pieczenik, Jerrold R Griggs, e Daniel J Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied mathematics*, 35(1):68–82, 1978.

[107] Vincent Moulton, Michael Zuker, Michael Steel, Robin Pointon, e David Penny. Metrics on RNA secondary structures. *Journal of Computational Biology*, 7(1-2):277–292, 2000.

[108] Michael Zuker e David Sankoff. RNA secondary structures and their prediction. *Bulletin of Mathematical Biology*, 46(4):591–621, 1984.

[109] S.B. Needleman e C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48(3):443–453, 1970.

[110] Ronny Lorenz, Stephan H Bernhart, Christian Hoener Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, e Ivo L Hofacker. ViennaRNA package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[111] Ignacio Tinoco, Olke C Uhlenbeck, e Mark D Levine. Estimation of secondary structure in ribonucleic acids. *Nature*, 230(5293):362–367, 1971.

[112] Jacques R Fresco, Bruce M Alberts, Paul Doty, et al. Some molecular details of the secondary structure of ribonucleic acid. *Nature*, 188:98–101, 1960.

[113] Charles Delisi e Donald M Crothers. Prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. USA*, 68(11):2682–2685, 1971.

[114] R. Nussinov e A.B. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci. USA*, 77(11):6309–6313, 1980.

[115] Ruth Nussinov e Ignacio Tinoco. Sequential folding of a messenger rna molecule. *Journal of molecular biology*, 151(3):519–533, 1981.

[116] S.R. Eddy. How do RNA folding algorithms work? *Nat. Biotechnol.*, 22(11):1457–1458, 2004.

[117] Richard Durbin, Sean R. Eddy, Anders Krogh, e Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge Univ. Press, 2000.

[118] Michael S Waterman e Temple F Smith. RNA secondary structure: A complete mathematical analysis. *Mathematical Biosciences*, 42(3-4):257–266, 1978.

[119] Ivo L. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Res.*, 31:3429–3431, 2003.

[120] Michael Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31(13):3406–3415, 2003.

[121] Ivo L Hofacker, Walter Fontana, Peter F Stadler, L Sebastian Bonhoeffer, Manfred Tacker, e Peter Schuster. Fast folding and comparison of rna secondary structures. *Monatshefte für Chemie/Chemical Monthly*, 125(2):167–188, 1994.

[122] Andreas R Gruber, Ronny Lorenz, Stephan H Bernhart, Richard Neuböck, e Ivo L Hofacker. The vienna rna websuite. *Nucleic acids research*, 36(suppl_2):W70–W74, 2008.

[123] David H Mathews, Matthew D Disney, Jessica L Childs, Susan J Schroeder, Michael Zuker, e Douglas H Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America*, 101(19):7287–7292, 2004.

[124] Ivo L Hofacker, Martin Fekete, e Peter F Stadler. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, 319(5):1059–1066, 2002.

[125] Stefan Wuchty, Walter Fontana, Ivo L Hofacker, Peter Schuster, et al. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2):145–165, 1999.

[126] Arthur L Williams e Ignacio Tinoco. A dynamic programming algorithm for finding alternative rna secondary structure. *Nucleic Acids Research*, 14(1):299–315, 1986.

[127] Norman Arnheim e Peter Calabrese. Understanding what determines the frequency and pattern of human germline mutations. *Nature Reviews Genetics*, 10(7):478–488, 2009.

[128] Lucieli Ceolin, Mirian Romitti, Débora Rodrigues Siqueira, Carla Vaz Ferreira, Jessica Oliboni Scapineli, Beatriz Assis-Brazil, Rodolfo Vieira Maximiano, Tauanne Dias Amarante, Miriam Celi de Souza Nunes, Gerald Weber, e Ana Luiza Maia. Effect of $3'$ UTR RET variants on RET mRNA secondary structure and disease presentation in medullary thyroid carcinoma. *PloS One*, 11(2):e0147840, 2016.

[129] Lucieli Ceolin, Débora R Siqueira, Mírian Romitti, Carla V Ferreira, e Ana Luiza Maia. Molecular basis of medullary thyroid carcinoma: the role of RET polymorphisms. *International Journal of Molecular Sciences*, 13(1):221–239, 2011.

[130] Steven R Morgan e Paul G Higgs. Evidence for kinetic effects in the folding of large RNA molecules. *J. Chem. Phys.*, 105(16):7152–7157, 1996.

[131] Jih-H Chen, Shu-Yun Le, Bruce Shapiro, Kathleen M Currey, e Jacob V Maizel. A computational procedure for assessing the significance of RNA secondary structure. *Computer Applications in the Biosciences: CABIOS*, 6(1):7–18, 1990.

[132] Shu-Yun Le, Jih-H Chen, e Jacob V Maizel. Prediction of alternative RNA secondary structures based on fluctuating thermodynamic parameters. *Nucleic Acids Research*, 21(9):2173–2178, 1993.

[133] Michael Zuker, John A Jaeger, e Douglas H Turner. A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison. *Nucleic Acids Res.*, 19(10):2707–2714, 1991.

[134] Mirela Andronescu, Anne Condon, Holger H. Hoos, David H. Mathews, e Kevin P. Murphy. Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics*, 23(13):i19–28, 2007. doi: 10.1093/bioinformatics/btm223. URL `http://bioinformatics.oxfordjournals.org/cgi/content/abstract/23/13/i19`.

[135] Geoffrey FS Lim, Gregory E Merz, Michael D McCann, Jocelyn M Gruskiewicz, e Martin J Serra. Stability of single-nucleotide bulge loops embedded in a GAAA RNA hairpin stem. *RNA*, 18(4):807–814, 2012.

[136] Brent M Znosko, Sara B Silvestri, Heather Volkman, Bob Boswell, e Martin J Serra. Thermodynamic parameters for an expanded nearest-neighbor model for the formation of RNA duplexes with single nucleotide bulges. *Biochemistry*, 41(33):10406–10417, 2002.

[137] Chuong B Do, Daniel A Woods, e Serafim Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90–e98, 2006.